Valid, Sensitive and Interpretable

A Novel Approach to EEG Analysis

Thesis

presented to the Faculty of Arts

of

the University of Zurich

for the degree of Doctor of Philosophy

By

Armand Mensen

from the Netherlands

Accepted in the spring semester of 2012 on the recommendation of

Prof. Dr. Lutz Jäncke and PD Dr. Ramin Khatami

2012

Table of Contents

Abstract **1**

Preface **4**

Why do statistics matter? 6

Chapter 1 - Current Methods of Statistical Analysis **14**

1.1 Conventional Analysis 16

1.2 Microstate Analysis 18

1.2.1 Analytical Approach 19

1.2.2 Evaluation 20

1.3 Multiple Comparisons Problem 22

1.4 Data Structure Considerations 22

1.5 SPM 23

1.6 Non-Parametric Solutions 26

1.6.1 Initial Calculation 26

1.6.2 Permutation Statistics 28

1.6.3 The maximum-statistic approach 31

1.6.4 Cluster Size 32

1.6.5 Cluster Mass 34

1.6.6 The Threshold Problem 34

1.7 Chapter Conclusion 36

Chapter 2 - TFCE **38**

2.1 General principles 38

2.2 The program 41

2.2.1 Inputs....42
2.2.2 Result viewer....44
2.3 Dipole simulation and sources overview....46
2.4 Signal Detection Theory....50
2.4.1 Definition of a true signal....51
2.4.2 Binary versus continuous classifiers....52
2.4.3 Limitations of signal detection measures....56
2.5 Optimal Values of E and H....57
2.5.1 Simulation Results....58
2.5.2 The ideal weighting for E and H....61
2.6 The Effect of Filtering....64
2.7 Chapter Summary....66
Chapter 3 - Direct Comparison to Other Methods....68
3.1 Previous Work on Method Comparisons....68
3.2 Simulated Source Comparison....71
3.2.1 Source data....71
3.2.2 The results....72
3.3 Direct comparison to SPM and GMD....76
3.4 Discussion on simulation results....79
3.5 Real data from SPM....81
3.5.1 Data source....81
3.5.2 Results....82
3.6 Real Frequency Analysis from SPM....85
3.6.1 Data source....85

3.6.2 Results 86
3.7 Reanalysis of previously published group data 86
3.7.1 Data source 86
3.7.2 Results 86
3.8 Discussion on real data results 88
Chapter 4 - Expansion to Complex Designs 93
4.1 Considerations for complex designs 93
4.2 Posner Paradigm 97
4.3 Method 99
4.3.1 Behavioural Task 99
4.3.2 EEG Recording and Analysis 100
4.4 Results 102
4.4.1 Behavioural: Reaction Times 102
4.4.2 Behavioural: Accuracy 103
4.4.3 ERP: Target Side 104
4.4.4 ERP: Cue Location vs. Target Side for SOA of 500ms 106
4.4.5 ERP: SOA vs Trial Validity 108
4.5 Discussion 109
Chapter 5 - Conclusion and Future Perspectives 110
5.1 Valid 110
5.2 Sensitive 111
5.3 Interpretable 112
5.4 Methodological Limitations 114
5.4.1 Bipolar deflections of a single source 114

5.4.2 Reference dependent 115
5.4.3 Recording parameters influence cluster sizes 117
5.5 Future work 118
5.5.1 Initial-statistics 118
5.5.2 Expansion of designs 120
5.5.3 Data smoothing 121
5.5.4 Nonstationarity 122
5.5.5 Optimal signal detection assessment 123
5.5.6 Software development 124
Chapter 6 - Appendix 125
6.1 How many permutations are sufficient? 125
6.2 Analysis method pseudo-code 126
6.2.1 TFCE calculation pseudo-code 128
6.3 Calculating neighbours 128
6.3.1 Triangulation of Electrode Coordinates 130
6.4 Analysis of source reconstructed data 132
Acknowledgments 134
References 136

Abstract

Electroencephalography (EEG) is able to measure brain activity on the microsecond scale and with an increase in the number of channels recorded can provide good spatial resolution. EEG has a long history in both clinical and experimental settings and has provided invaluable information on the brain's functioning. Recent advances in the capabilities of EEG recordings have made the issue of how to statistically tackle the large datasets unavoidable.

Threshold-free cluster enhancement (TFCE) has recently been shown to be a superior technique in the analysis of fMRI datasets. Combined with non-parametric permutation based statistics, this thesis shows that TFCE can also be applied to analyse EEG datasets. TFCE essentially finds clusters in the data over multiple thresholds and combines the information with the strength of the signals in that cluster, enhancing weak but clustered signals to a level directly comparable to strong focal signals.

Chapter 1 provides an overview of the variety of methods currently available. This includes the conventional analysis techniques, microstate analysis, statistical parametric mapping, and permutation approaches using intensity and cluster based statistics. A particular emphasis is placed on how, in one way or another, they have failed to become the standardised method sought after. This is in spite of the fact that most methods have been available for quite some time.

The second chapter makes a formal presentation of the TFCE method starting with the fundamental reasoning behind the approach which maximises both statistically validity and signal sensitivity to a wide range of signal types. The resulting output of

the method is then explored; along with the result visualisation program to show how the properties of the analysis lead to a maximally interpretable outcome. Subsequently, ideal weighting parameters for the analysis are found both theoretically and empirically using a broad range of simulated source datasets.

Chapter 3 then deals with a direct comparison to the methods discussed in the first chapter using those same simulated sources, as well as three diverse datasets from real experimental settings. Here it is shown that the TFCE method, with both its theoretical and empirically derived settings, generally performs better and more consistently than all other methods tested in terms of sensitivity while still maintaining strict control over the number of false positives.

In chapter 4 the method is expanded to more complex experimental designs common to the research field. Initially the auxiliary considerations are discussed from a purely theoretical perspective and then put into action by analysing results from a complex experimental design on the orientation of visual attention. In the final chapter, a summary of the work is given along with possible future research that could be done to further enhance the methods capabilities.

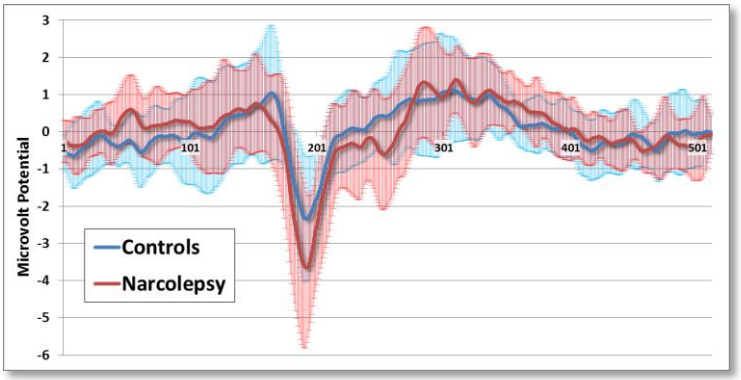
The purpose of this thesis is three-fold. The first is to make the reader aware of the various difficulties in EEG analysis and the advantages and disadvantages of previously proposed solutions. The second is to present the TFCE method as a viable alternative; and act as a guide for researchers who have faced similar issues and asked the same questions. Thirdly, to demonstrate the superiority of the TFCE method in comparison to other methods but nonetheless highlight areas where it could be improved in future work.

For my parents...

Preface

The problem is misleadingly simple to state: given two or more event-related potentials (ERPs) in electroencephalography (EEG); how are these waveforms different from one another? The two ERPs might be generated from different groups of people conducting the same experiment such as men versus women, or patients vs controls; or they may be two different conditions in the same experiment, such as the brain's response to happy versus sad pictures, or auditory responses to standard and rare frequencies. The ERPs of various experiments are usually created using fairly standardised procedures, and tend to all look fairly similar to the untrained eye. Figure 1 shows an example channel from a real dataset which compared patients with narcolepsy-cataplexy on a motor inhibition task. However, even here with a single channel example there are potential differences over multiple time points with different ranges and peaks. Given the variance in these average waveforms for both groups, it is impossible to determine statistical differences without an advanced method of analysis.

Figure 1 – Real example of a single channel ERP.



To be certain; the problem stated is a large one in terms of both importance and sheer scale of the datasets. Modern ERP datasets usually consist of a large number of recorded channels (up to 256), with a broad time range around the event in question and a high time sampling rate (usually no less than 100 samples). These datasets are very comparable in size to those of other neuroimaging methodologies such as functional magnetic resonance imaging (fMRI), and the questions posed are similar. The EEG research field is still growing and thousands of articles are published each year in peer-reviewed journals[*]. The creation of ERPs, without a doubt, forms a large part of that research. However, despite more than 100 years of research in the field of EEG, no method has yet to step up and reliably answer the proposed question.

This thesis aims to break down the problem into several parts, and propose a novel method which solves these issues step-by-step. As the title of this thesis suggests, the solution must be valid, sensitive, and interpretable. That is, valid from a statistical perspective in that we make little to no compromises in the data's integrity, eliminate sources of bias, and control for risk of making a false positive prediction. That is, stating a difference is there when it really isn't. A method's sensitivity is how well it is able to detect a difference in signals when there really is one there. Too often sensitivity is seen as a direct trade-off to a method's validity; and researchers will often take measures which increase sensitivity at the risk of having ultimately false research findings. Although the trade-off is true in some aspects, the relationship is no means direct or linear and there are ways to optimise both concurrently. A method's interpretability on the other hand is more difficult to define; yet there are a few principle aspects which would likely be

[*] 4935 since the start of 2011 alone (according to a PubMed search).

agreed upon. A method is interpretable if: the results can be understood without having to understand the precise details of the analysis; the structure of the results is directly comparable to the structure of the original data; precise claims can be made about differences in specific channels or time points; and finally, the big picture can be easily understood and summarised.

Why do statistics matter?

To a large extent, this thesis discusses a statistical procedure. One that should ultimately help decide how strongly one should believe in differences found between ERP waveforms. Given the common apprehension to discuss any statistical process, there are two relevant points that should be made before any specifics of the methods at hand are dealt with.

Firstly, although the statistical process is founded in mathematical theory, it need not be described with it. Statistics should be seen as a formal translation of our intuition about the data. For example; in order to quantify the real difference between any two sets of numbers we are tempted to simply look at the difference between the averages of those two groups. If we see however that those numbers vary a considerable amount, our intuition will already have lowered our belief that the simple difference in averages carries any importance. The commonly used t-statistic is nothing more than the differences in means, normalised by the variation of the two groups. Thus, the t-statistic can and should simply be seen as a formal description of that mental calculation that has already taken place (albeit perhaps somewhat unconsciously). The problem however is that with very large sets of data, like the ones we currently face in EEG, our intuition fails at seeing the whole picture; and we must rely on the statistical

process to convert that data into useful information. That said, each part of that statistical process should be simple enough to understand intuitively. If a process no longer conforms to our own intuition we should be encouraged to examine it fairly critically. The statistical analysis of EEG slowly has come to the point where relying on intuitive data description becomes problematic. At this point, the research community became divided in that researchers either chose to continue with simple analysis methods at the cost of bias, specificity, and data integrity; or attempt to confront the growing complexity at the cost of sensitivity. The current thesis aims to bridge the gap by presenting a method which has learned from the previous analysis attempts and handles the complexity in the data using directly intuitive ideas.

Although most of thesis deals with the first general point about statistics, there is a highly relevant second point which demands discussion. Essentially, statistics matter because they are far too often misunderstood and used incorrectly. Thus is because statistics are simply a set of tools to describe numbers. The tool's fundamental purpose is to turn data into information. Yet being a tool, statistics may be inaccurately used and the information extracted biased, manipulated, or incomplete. As several critiques have pointed out, this error is common to all scientific research. In a review article Ioannidis (1), made the bold statement that most research claims are false. In his review distinct points are made to support his statement and it seems as though EEG analysis has been particularly vulnerable to many of these.

Firstly, an increase in the number of variables tested will lead to false findings. This is also known as the multiple comparisons problem and will be dealt with in more detail when discussing currently available methods in section 1.3 Briefly, with each test conducted, one increases the chances that a completely

random result will show a sufficiently high deviation and be proclaimed true. With EEG data, not only are the groups or conditional factors independent measures, we may also consider each channel and time point analysed as an independent test.

A second point is that with an increase in the flexibility of experimental design, outcome measures, and analytical techniques, the odds are the published findings are false. Each level of flexibility will introduce a new line of analysis which could ultimately turn a negative result into a positive one. This results in the same issues as multiple measures, but the tests are conducted on the same original data. In EEG research flexibility can, and does occur at multiple stages. During the recording stage, the experimenter has an array of choices for number of electrodes and configuration. Due to the relative ease of setup, there are few hindrances to the overall experimental design that can be currently conducted. During the pre-processing stage, there are few standards to the type of frequency filters to be used, as well as the multitude of artefact correction techniques.

Despite the lack of standards for the two stages just mentioned, there can be no doubt that it is the flexibility of outcome measures and analysis techniques that dramatically increase the probability of producing false research in the EEG field. There seems to an endless amount of information one can extract from the raw data; simple microvolt value, latency to maximal peak in the wave, frequency power, source localisation, connectivity, amongst several other possibilities. Once we have decided on some outcome measure (or range thereof), we can bombard it with an armoury of available statistical techniques. As we shall see in the next chapters, the history of EEG is long enough that some theoretical justification for the choices can certainly be found in some previous research. With so much flexibility, the journey from

raw data to a significant result seems to be more a matter of patience and perseverance rather than on the truth behind the data.

The third relevant argument states that the more independent research teams there are involved in a field and the hotter that field, the more false research reports tend to appear. This point requires further explanation as it seems to contrast to the common perception that the more research that is conducted in a field, the closer we will get to the truth. The key here is that research teams are all too often independent, with no sharing of data or results. Furthermore, with the all the flexibility involved in EEG research, they do not often produce directly comparable results, even about the same research question. This in turn creates another multiple comparisons problem; here however the multiplicity comes from the various groups. Thus, one of the many research groups is bound to find some significant results among what could very well be completely random data. The hotter the research field, the larger the pressure will be to publish even small significant results and ignore the overall negative results. Over the last decade the hype of functional magnetic resonance imaging (fMRI) research is beginning to settle, and the relative low cost and ease of use of EEG is making it a popular investigative method once again.

Given just how well the arguments for false research apply to the EEG field, one should certainly begin to worry about the quality and sincerity of many of the published reports. In fact, given the propensity for false research in the field; the size of the claims made by researchers may be regarded as proportional to the amount of bias in that study. It is important to note that much of this bias is completely unintentional, and this critique is not directed to the research standards of the individual experimenter. Rather,

the bias is more the result of the methodologies available in the field as a whole.

The previous issues with research findings make the reported significance value, the p-value, untrustworthy. However, the arguments at least assume that that the p-value is still calculated correctly and the ability to interpret it is intact. However, several reviews over the last decade bring even these assumptions into question.

It is common practice, alongside the p-value, to report both the actual statistical value (e.g. t-value or F-value), along with the degrees of freedom. This should in fact be a redundant procedure because any one of the terms may be calculated when the other two are known. However, all the papers published in the highly respected, and peer-reviewed journals Nature and British Medical Journal in 2001 where checked precisely for this redundant fact (2). Surprisingly, the authors of the review found that over 11% of the reported p-values results in both journals were incongruent with the reported statistic and degrees of freedom. It is even more surprising when you consider that this value could only calculated where exact p-values were given, as opposed to the unfortunately common reporting of the p-values being over or under a given threshold (e.g. $p < 0.05$). It is interesting to note that a later review of this article found that those authors used an invalid statistical method to come to some of their other conclusions (3). It is not unreasonable to assume that if the research submitted and published in these two journals is guilty of what is a very simple statistical mistake to check, then this percentage is likely to be higher in many other journals. More recent reviews have confirmed this suspicion, and also indicate that the mistakes made are more often than not in the researcher's favour (4–6).

A much more difficult matter to dissect is how researchers understand and interpret the p-values they calculate. At its core, the p-value is an indication of how likely it is that a more extreme statistical value would be found if the value was indeed due to random chance. As most researchers interpret it, I would say correctly, the p-value is an indication of how much trust one should have in differences found between datasets. I would not be the first to point out that the significance tests are not constructed to provide binary outcomes of yes and no but to give an indication of the chance of error along a spectrum (7). Although, modern thinking has led to a standard threshold of 0.05 being the cut-off point for statistical significance, this is ultimately misleading. As Fischer, the pioneer of significance tests, pointed out:

If P is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05. . . .”(8)

Thus it is the researcher who must decide how to interpret that chance value in light of previous findings, the quality of the methods used, the actual size of the effect found and the potential impact of the finding.

This fact is not always clear, even among statisticians. In a recent review of significance tests, Hayat seems to get the importance of p-values completely backwards (9). He argues that since p-values of 0.04 and 0.0001 fall below the standard 0.05, they should be regarded as equivalent evidence against the null hypothesis and no further conclusions should be drawn. Also explicitly stated is that p-values of 0.06 and 0.04 should warrant completely different conclusions since they lie on opposite borders of the 0.05 value set. It is precisely this view that leads to the poor practice of reporting significance with the symbols ‘<’ and ‘>’, rather

than simply quoting the value, which takes precisely the same amount of space on a page and effort on the part of the experimenter. The view that 0.05 is a strict cut-off with some inherit and mystical property to find truth is a common one, but ultimately the product of misunderstanding and a mind that prefers black and white answers over gray areas.

In rather crude terms, significance values should be inversely proportional to how much money you would want to bet on the result found being true. In this way, p-values of 0.04 and 0.06 should warrant approximately the same wager, whereas you should consider raising the stakes when confronted with a p-value of 0.0001. As we shall see in the first chapter, many of the current methods of EEG analysis rely on extra quantification of the p-value found. This is done by specification of its original statistical value, or its surroundings, or its experimental importance. This further quantification of the statistical results found can ultimately lead to any unexpected 'significant' effect being explained away, or any non-significant effect being accepted as a hypothesis. As already argued, p-values are not the only measure that should be reported; however, the actual value discovered by a test should be a useful one that needs no further special explanations to be understood.

If p-values give an indication of the chance of being wrong when you think you are right, then power calculations give us an indication of the chance of us being right when we say we are wrong. This is the other side of the coin of statistics that is largely omitted from explicit calculation. This is somewhat understandable given that in science, with the large exception of medicine, we pay a much higher price for false positives then we do for false negatives. Yet, we cannot ignore this area of statistics so readily since it is, after all, a scientists job to find out the truth about how to world works, and not simply to be a cynic. Strictly speaking, there is

always a trade-off between these two aspects. We increase the risk of committing one type of error in order to reduce the risk of the other. Many of the differences in current methods of EEG analysis are simply variations of where along this spectrum the lie. Most methods in daily use have largely sacrificed their statistical integrity and biased their statistics towards the goal of increasing their chances of finding a positive result. Given limits of current funding, competition for jobs and results, publication pressure, and the time and energy invested in research, this bias is to some extent understandable.

Although understandable, it is undoubtedly damaging in the long run. There is a reason why the history of science has been far more careful with false positives than false negatives. A single positive finding can lead to multiple groups trying to replicate the finding, expanding on the finding, or using the result to spur more exploratory studies in the field. This takes an enormous amount of funding, time, and organisation. Yet, if there is good reason to believe something is true despite a negative result, there is bound to be some group that carefully examines the finding again.

In summary, proper statistics are important because:

- our eyes can often not see the whole picture (especially one so complex as the results of EEG experiments)
- we are highly susceptible to many sources of bias
- misunderstanding has often lead to incorrect application
- the financial cost is so high when we get it wrong
- and ultimately, mistakes in this area are so damaging to our pursuit of truth

Chapter 1 - Current Methods of Statistical Analysis

Once individual ERPs have been obtained from the raw data there seems to be an endless amount of options in order to statistically compare them [*]. Having several options is not necessarily a negative point as long as there are certain reasons to choose one method over another. The reason could be by experimental design, data distribution or specific statistical outcome. However, in the EEG field, no such decision tree exists and yet the statistical goal is almost always to simply localise significant ERP differences in time and space. Methods of choice seem to better correlated with the research groups involved, or the specific field of interest rather than with the characteristics of the data or the experimental hypotheses.

Table 1 presents a literature overview of the most recent studies at the time of writing and just how varied analysis techniques can be. Six of the ten studies presented use only a fraction of the channels they recorded in their analysis and the remaining studies only use the information to calculate some spatial average or other summarising statistic. Moreover, eight of the ten studies do not look at the all the time points in the ERP and select, mainly without empirical justification, only some of interest. Even of the two that look at all the time points in their ERP, the information is averaged over 'windows' of time.

[*] Although the pre-processing stages of filtering, artefact correction, and averaging are crucial steps in determining the shape of the data taken for further statistical analysis; they are explicitly not discussed here. This is not only for space consideration but because the methods for pre-processing are already fairly well established and there is general consensus about the optimal methods.

Table 1 – The ten most recent studies found using the search terms "EEG" and "ERP" in PubMed. The table shows the article reference, ratio of channels recorded and subsequently analysed and the methods used. Note the relatively low ratios of channels recorded and actually analysed; the different variables extracted from the data; and the many different statistical methods used to answer the same basic questions.

Ref	Channels Tested / Recorded	Variable	Statistics
(10)	64 / 64	Amplitude Differences at P30, N45, P40, P60, N100, N120, P180, N280, N400, P1000	MaxT Permutation
(11)	2 / 62	Amplitude and Latency of N200 (Fz) and P300 (Pz)	ANOVA (no correction)
(12)	1 / 32	Amplitude of Peak to Peak N2-P2 component (Cz)	ANOVA
(13)	64 / 64	Partial Least Squares between conditions	Not described
(14)	60 / 60	Amplitude averaged over 25ms time windows	ANOVA for each time window (3 window correction)
(15)	24 / 64	Average amplitudes over three time windows (300-2000ms) and electrode regions.	Separate ANOVA for each region (no correction), then t-tests
(16)	3 / 62	Amplitudes over frontal channels	One-way T-tests to baseline
(17)	2 / 64	Average in frequency and time domains for user selected maximum channel	Multiple ANOVA tests (no correction)
(18)	3 / 128	Amplitude and latency of components N20, P25, N35	Mann-Whitney non-parametric test
(19)	18 / 18	Principle component analysis of N1, P2, N2 and P3 / Resting frequency	Pearsons correlation of resting state frequency and components

1.1 Conventional Analysis

Much of the current statistical analysis of different EEG samples has attempted to simply avoid the increasing complexity of modern EEG datasets (20). The goal of various methods has been to reduce the data's inherent complexity: often, only certain channels are taken over specific time points (21–23); or channels are grouped into areas such as left and right hemisphere (24); or samples are averaged over chosen time windows. The few measurements left for analysis are then subjected to simple t-tests or in more complex experimental designs to multiple but independent analyses of variance (ANOVAs).

There are multiple issues with any of these methods. Firstly, it is wasteful of the data collected. Data which required increased costs of equipment to measure more channels, increased time for subject preparation, and increased computational resources to record and analyse. Secondly, it usually involves various levels of user interference and, more often than not, arbitrary selection. This corruption of statistical validity is most often justified with an appeal to a-priori selection based on previous literature and established pathophysiological considerations. Although in limited cases this may be rationalised, this approach is too open to user biases (25) and with an abundance of literature in the EEG field one may find reason to pick out any number of channels or time points.

Furthermore, one should also be able to show that while some measurements show significant differences; these stand out amongst neighbouring channels or time points which are not significant (26). For example, if a study found that two groups of participants differed in a certain frequency band, this may only really be theoretically relevant when neighbouring frequencies are

in turn not significantly different. Sacrificing neighbouring frequencies or neighbouring channels for purpose of complexity reduction will ignore this information. Thus although the test may be more sensitive to the desired effects; the specificity of the results is left unknown. Similarly, one may be interested in the boundaries of an effect, and not just some maximum point of difference (27) (e.g. when do the differences first emerge, rather than when they peak). Secondly, with ever increasing complexity in paradigms, experimental manipulations, and novel ideas, one is highly unlikely to have sufficient specific evidence to justify the user choices inevitably related to the data reduction. And in doing so, might completely miss on unexpected effects.

Even in the cases where there is a very specific a-priori reason for the researcher's selection there is still good reason to analyse the whole data as an additional measure. Reporting only the results of the channels or time points corresponding to the specific hypothesis constricts the data's usefulness to other members of the scientific community who may have other hypotheses about other time points or channels under the same experimental conditions.

A more data-driven process of complexity reduction has been to use principal component analysis or independent component analyses on the raw data then select only a few of the temporal and spatial components for further analysis (28, 29). These techniques depend on finding components which explain most of the variance, however given the low signal-to-noise ratios (SNR) of ERP signals in the raw data, it is unlikely that the principal components found necessary correspond to task-relevant events (30). Thus, results tend to be unstable and still require a large number of input parameters and assumptions about the data. Furthermore, when significant differences are found, it is impossible

to directly trace these differences back to the original channel-sample pairs of interest.

1.2 Microstate Analysis

Murray provides a comprehensive review on EEG analysis using the theoretical underpinnings of 'microstates' (31). The term microstate comes from the initial observation that large scale changes in the states of consciousness of the brain (namely, wake and sleep), are accompanied by large scale changes in the shape and topography of the measured EEG waves. Taken a step further, we may also observe that in a much shorter time frame of a typical ERP, there also seems to be periods of stability in the EEG topography followed by rapid changes to another configuration. Thus, a period of stability, even if brief, can be termed a microstate since if there is no change in topography, we can assume the same neuronal generators have remained active. More accurately stated, different map configuration must have been caused by different generators since it is theoretically possible that different generators could cause the same topography, just as different objects may cast the same shadow.

Upon reading the introductory paragraphs of Murray's review of the procedure (31), one can see that the arguments made against current EEG analysis methods are virtually interchangeable with the ones made in this thesis. The vast size of datasets, user biases, and wasted data by a priori selection, are all discussed with critical examples. However, the suggestions on how to overcome these issues quickly diverge from there on. This begs the question, how can the same path lead to different destinations, and can we make some conclusions on which method is more appropriate in given situations or prior assumptions.

1.2.1 Analytical Approach

Classifying each time point as a certain microstate involves using one of two pattern recognition algorithms; *K-clustering* or an *atomize and agglomerate hierarchical cluster analysis*. Although each algorithm has their own strengths and weaknesses, both essentially identify ideal topographic maps over the grand averaged data which taken together describe an optimal amount of variance (optimal in that a high amount of variance can be described with a minimal number of maps). These 'template maps' can then be put back on the ERP time series in order to visualise the different microstates apparent in the ERPs. Importantly, the distinct microstates assigned to the averaged ERPs should not be necessarily viewed as *significantly* different since the templates are calculated from the data and then merely applied back to the data. Thus, there is always an assigned microstate, even if it does not fit back on the data *well*. One form of statistical significance can be calculated by fitting the template maps back on the individual datasets, and performing analyses on how well this fitting procedure describes each dataset. Authors proposing the microstate analysis have suggested several different parameters which can calculate and analyse the 'goodness of fit' but are reluctant to promote one measure over the other and each has its caveats. Instead, two quite different measures are generally advocated to assess the statistical significance level for the differences between two ERPs (32). Crucially these measures are completely independent of the microstate fitting procedure.

The first measure is the relatively common Global Field Power (GFP), of a signal. GFP can be roughly defined as the total squared deviation of each electrode, then normalised by the number of electrodes. When looking at the GFP, we are informed about the strength of activity in the brain at each particular time

point. The drawback of course is that we have lost information about the topography of the ERP, and identical GFP values can be generated from completely different topographies (e.g. exact opposite topographies).

The second measure constructed is therefore an attempt to summarise how different two ERP topographies are (even when GFP is similar). Global Map Dissimilarity (GMD, DISS or Global Dissimilarity Index (GDI), depending on the article), is essentially a measure of the spatial correlation between any two maps (either different conditions, or time points in a given condition), normalised by the instantaneous GFP. Importantly, neither of these two measures constitute a statistical test as they stand. Since the measure of GMD is a single measure, without means and standard deviations, parametric statistics cannot be conducted. Therefore, a non-parametric permutation test is necessitated (see the later section 1.6 for a description of permutation tests); oddly enough named a 'topographic ANOVA'. The permutation test run however, does not seem to fairly take account multiple comparisons in that new empirical distribution is calculated for every sample rather than a single distribution (made from maximal values), over all samples. In practice, a somewhat arbitrary correction is made in that a minimum consecutive number of significant samples must be attained in order to justify further inference (33–35) It should be noted that both these measures are calculated on data which has been average referenced.

1.2.2 Evaluation

From a purely statistical perspective, the process described by proponents of microstate analysis is little more than performing two separate permutation tests which are uncorrected for multiple comparisons across samples. The two measures, GFP and GMD,

emphasize two different aspects of the data but both essentially come from a single measure of uV in the ERP, and can thus be regarded as (relatively good), attempts to reduce the complexity of the data in different ways. If we include the additional measures of the microstate fitting procedure we are confronted with a wholly new multiple comparisons problem as we are analysing several tests on different measures which in the end still come from a single measurement of uV in the data.

It is also crucial to realise that the measures of GFP and GMD, as well as those of the microstate analysis are not dependent on one another in any way. For the most part, the microstate analysis is not one of statistical relevance. Like that of source estimation, the analysis will always produce a result; which would then still need to be further confirmed using statistical significance tests. In conclusion therefore, microstate analysis should be viewed as a potentially very useful tool in visualising the data, which could then be used to aid interpretation of the results. Therefore, if the user is willing to accept the theoretical underpinnings of the microstate procedure, there is no reason why the analysis cannot be run once an appropriate statistical test has determined significant differences in the original datasets.

Given that each method of reducing the complexity in the data has its own computational costs and validity drawbacks, what is needed is a data-driven process which uses all the data collected, with little to no input requirements from the experimenter, in order to maintain statistical integrity while being sensitive to the various possible differences between EEG signals.

1.3 Multiple Comparisons Problem

Large datasets imply a large number of comparisons between channels, time points and groups. With each comparison comes the risk that any differences found are purely by chance, the type 1 error. This is the 'multiple comparisons problem' which has been the topic of years of discussion in the science community (36). As discussed in the previous section, conventional analysis has attempted to avoid the issue altogether by only testing a small fraction of the data. A common approach to correcting for this bias with a small number of comparisons is the Bonferonni correction; here the significance threshold is lowered proportional to the number of tests being conducted. However, with 256 channels and 500 time points, p-values would have to be in the order of 10^{-7} in order to be deemed significant. If we were forced to take such a stringent approach to our data, we would never publish a positive result. The Bonferonni correction however assumes that each data point is independent of one another, whereas in EEG information from near sensors and time-points tend to be highly correlated. This fact has allowed the construction of several less conservative correction methods which nonetheless provide a strong and valid control the risk of false-positives to acceptable levels.

1.4 Data Structure Considerations

The remaining approaches described here were principally developed with fMRI data in mind and applying them to EEG requires specific considerations. Primarily the data structure is different, with fMRI data being three dimensional, where each dimension represents the same type of relationships as the other, namely location in space. Furthermore, the total size of the dimensions is well defined by the scanner resolution, as well as the

relationship between points being a constant value measurable in millimetres. For EEG data, the data structure could be two or three dimensional with channel, time, and frequency as possible dimensions of the data. Importantly, the size of the data structure is highly variable between different datasets. Different numbers of channels could have been recorded, and even if the same number of channels were recorded their positions could be drastically different from one dataset to the next. Moreover, both the number of samples or frequency bins is not defined from the recording and will depend on user preferences or the frequency bins of interest. However, in many cases, EEG datasets tend to be in the same order of magnitude of size as those in fMRI experiments, and thus similar statistical methods should be expected in dealing with the large number of data points.

1.5 SPM

Statistical parametric mapping (SPM) is a software package which works as a toolbox in Matlab (Mathworks, Inc.). It was designed as a tool for the processing and analysis of neuroimaging data from PET and MRI, and has rapidly become the leading analysis tool in those fields. In the meantime several different versions have been development and countless extensions have become available. Not long after its development, two papers were published to show how the software could also be used for the analysis of EEG datasets with the same underlying principles of the other imaging modalities (37, 38). This is an attractive property for several reasons. Firstly, measurements techniques are increasingly being used in combination with one another. The simultaneous combination of EEG and fMRI can be used to locate activity in the brain with high precision in both the temporal and spatial domain (39–43). Furthermore, the combination of EEG and TMS has allowed

us to influence brain activity without having to rely on sensory stimulation (44–48). Secondly, field-specific biases in analysis would tend to be minimized. Lastly, it would allow for more direct comparisons of results across studies, even those which used different investigative techniques in their design.

In SPM, the ERPs may be analysed in two forms. One option is to conduct an appropriate source analysis and the scalp data becomes a 4D, whole brain, time-series. Or, the original surface information is projected and interpolated to create a flat 2D image, and images are stacked over time to produce a 3D image of interpolated channel-sample pairs. There are issues with both of these methods but these are discussed later in section 6.3 SPM is a mass univariate approach, which essential implies that, at the first stage of calculation, each measurement (channels, time points, participant), is taken as a separate measure and any correlation in the data is not included in the model. For SPM the covariance in the data is accounted for at the statistical-inference stage by adjusting significance values using random field theory (RFT). As an oversimplification, RFT attempts to find a statistical cut-off threshold based on an estimate of the number of independent elements in a smooth image given certain assumptions (see (49) for a comprehensive overview).

Several further methodological papers have since been published detailing additional procedures of SPM for EEG data (50–55). Given its popularity in the MRI field, and the sound theoretical background that that has brought, is may seem fairly surprising that SPM has not become the standard analysis technique in the EEG field. In fact, SPM could only be found to be the principle analysis

method in 6* publications to date(56–61). Looking at how SPM was used in those studies may give us a reason. In none of these studies was SPM used exclusively, or strictly as presented in the guidelines. Two studies (unnecessarily), only analysed discrete time windows (59) and one subsequently based inferences only on the uncorrected results (56) (thus ignoring the overall purpose of the SPM procedure). Another study, which used a combined fMRI-EEG approach actually just used paired-t comparisons and the specific SPM results could not be found (61). Two studies by the same group seemed to use SPM methodology correctly but then repeated essentially the same analysis using traditional ANOVA statistics in order to provide channel specific effects (57, 58). Lastly, one article only used SPM as a secondary analysis procedure while classical ERP component amplitude and latency analyses were presented in greater detail; and no explicit confirmatory comparison was made between the two results (60). Furthermore, none of the studies actually presented the SPM results tables and figures readily available after analysis. Thus, it may be more telling to ask why they chose to use SPM at all, and a quick examination of the authors, or acknowledgements reveal many familiar names to the fMRI field.

In conclusion, even a rigorous analysis technique that SPM is seems to still be quite open to user biases in analysis considering each paper applied the approach in a slightly different way. The presentation of the results also seems to be unintuitive since it is automatically generated and yet rarely used in description. Additionally, due to its design having been built for PET and MRI datasets, the analysis process can be rather complex and unintuitive

* Using a pubmed search for the terms “EEG” and “SPM” as well as reviewing the articles which cited the original article by Kiebel and Friston.

to undergo[*]. Finally, it may be the case that potential users are unsettled by the multiple assumptions that underlie the SPM procedure and whether they are appropriate to apply to their EEG datasets.

1.6 Non-Parametric Solutions

Non-parametric permutation statistics require relatively few assumptions about data structure in comparison to its parametric counterparts. Parametric assumptions, such as normality, homogeneity of variance or sphericity, become hard to attain with increasing number of variables. More importantly here however is that non-parametric based statistics, such as the permutation statistic described here, offer an easy and intuitive way of dealing with the analysis of multiple sensors and time points while maintaining strict control over the false alarm rate. A further advantage is that any statistic can be used that one sees fit to describe the differences in signals. Permutation based methods have long been seen as ideal statistical tests but have largely been ignored as the computational cost has classically been too high (62, 63). However, with current computers and optimised algorithms, processing times do not differ significantly[†].

1.6.1 Initial Calculation

The first step is to calculate an initial-statistic which represents the difference of two signals at each channel and time

[*] Conducting the comparative analyses for this research was rather tedious, with very little confidence the correct procedure was being followed (even when analysing SPM's own tutorial data).

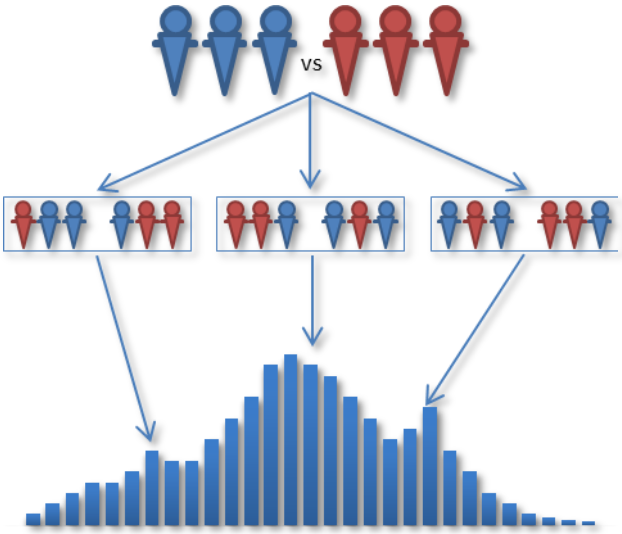
[†] For the analyses presented later in 3.5 the SPM method took 112 seconds, while the TFCE approach took 238 seconds. However, this SPM time does not include the necessary, and lengthy step of converting data to images.

point. Its selection does not affect the permutation process and can be chosen by the user to suite whichever type of difference is best suited to the experimental hypothesis. For instance, if absolute differences are the only concern, then the mean difference between the two samples can be used. On the other hand, if the actual values are irrelevant and differences in signal variance are the main concern then standard deviation or variance can be calculated. In most cases however, what is generally thought of when signals are described as 'different' is the difference in mean normalized by the variance; which is essentially the t-value. It is important to note here that although the t-statistic is generally associated with parametric testing, here we do not use the t-values directly to determine statistical significance (under the known t-distribution) and so do not need to make the same assumptions.

For more complex experimental setups, common to modern EEG research, with multiple groups or multiple factors, an F-test can be performed as in the typical analysis of variance (ANOVA). Here, each F-value, representing main effects and interactions, would be included in the permutations and would be viewed separately in the results. As with any ANOVA, if a group with more than three levels or an interaction is significant, multiple individual tests need to be carried out post-hoc to determine where the precise difference is (see Chapter 4 for further details on complex designs). For the primary tests in this thesis, the choice of initial-statistic was limited to either the independent (unpaired) t-test assuming unequal variances or the dependent (paired) t-test.

1.6.2 Permutation Statistics

Figure 2 - Basic illustration of the permutation approach. Here, two groups of 3 people can be permuted to give rise to several new groups each resulting in a new comparison statistic. All the statistics from the randomised group go into making a dataset-specific null distribution which is used to make inferences about how likely the original labelling actually is. This gives rise to an exact p-value for each dataset.



Since the choice of initial-statistic is open, and the number of comparisons to be made is highly variable, we have no frame of reference as to how large a difference should be before we can deem it to be significant. Unlike for parametric statistics, there is no pre-existing table which we can look up to determine where our significance cut-off point is. Under the permutation approach, a new frame of reference is calculated from the data itself which is specific to each analysis conducted. The permutation method works under the fundamental principle of: *if the null hypothesis of no experimental effect is true, then the labels we applied to our dataset are meaningless for our measurement*. Thus, a new dataset is created by randomly permuting the given labels such that some

members of the first signal are now part of the second signal and vice-versa. Hence, in any newly permuted dataset, there will be members of different conditions or groups now being part of the same label. This random exchange (permutation), of labels is done a sufficiently large number times to obtain new sets of initial-statistics each time with which to construct a new null distribution of data. The idea is that if the null hypothesis were true, the initial-statistic would about the same as the values found in randomly exchanged datasets. If the null hypothesis was false, and there were real differences between groups, we would expect the initial-statistic to be quite different from the other values obtained under permutation. The p-value of a given difference is the proportion of permuted statistics that are more extreme than the statistic from the original dataset. That is, the maximum possible p-value is dependent on the number of randomised datasets created. Importantly, the result of a permutation test is exact in that the p-value obtained is precisely equal to the rate of false positives in the long run.

For example, imagine our two signals came from the averaged ERPs of two groups of participants, patients and controls. If there were no difference between patients and controls for this experimental manipulation, then the labels are meaningless and a patient could have just as well acted as a control in this situation. So in constructing our null distribution, we can take all the participants of the study and assign each participant the label of patient or control randomly and calculate a new statistic for each channel and time point from this data. If we had a total of ten participants in each group then there would be ‘n-choose-k’ ways in which we could permute the labels, thus creating 184,756 distinct

datasets*. Fortunately it is not necessary to build every possible dataset in order to obtain near exact results (see section 6.1 of the appendix). Since the number of permuted datasets created is directly related to the time taken for the analysis to run, typically values between 1000 and 10'000 of random permutations have been used to build the null distribution. As the power of computing continues to increase (or the patience of the researcher), this number will continue to get closer to the real maximal value.

This procedure only needs slight adjustment in order to deal with the multiple comparisons problem in EEG datasets. Rather than build a null distribution for each channel and time point, a single null distribution is created under which all samples can be compared. This is achieved by taking only the maximum value from each new dataset for the empirical null distribution. Note that the maximum value could come from any channel and any time point for each permutation. Thus, the entire permuted dataset is summarized by a single value, herein called the *summary-statistic*, which then forms a single data point in the histogram of the null distribution. The exact p-values for each original channel and time point is then calculated as with the single variable case by finding the proportion of permuted datasets which show values more extreme than the statistic in question.

The permutation process will always result in a valid statistical test, in the sense that the false positive rate is controlled. The justification for this has been well explained elsewhere and will be left out here to maintain user readability (64). Although the validity of the statistical process is clearly important, it's also useless unless the type of differences we are interested in can be detected

* For one sample tests the number of permutations is 2^n. For correlation analysis the number is n-factorial (n!). For one-way ANOVA it is a generalisation of n-choose-k:

$$(n_1 + n_2 + \cdots + n_k)! \,/ (n_1 \times n_2 \times \ldots \times n_k)!$$

by the method. Furthermore, the end result should be of a structure that is interpretable, especially given the vast amount of data. Therefore the choice of summary-statistic which is calculated from the original data, and taken from each permutation to form the null distribution, is a crucial element in making a test sensitive. Here, we will thus consider the various suggestions presented in previous work and discuss issues and how the data structure of EEG affects each.

1.6.3 The maximum-statistic approach

Taking the absolute maximum comparison value from each permutation is the most basic summary-statistic that can be taken. Here, each newly created dataset is examined for its highest initial-statistical value and the rest are ignored. The permutation distribution is then built from these maximum statistics, one from each permuted dataset. Finally, each original channel-sample pair can be directly compared to this calculated distribution to determine its p-value. In fMRI research this method has been called the voxel intensity approach since the only criteria of importance is the intensity of each data point, and not its location in space.

This is the technique that is usually implied when studies have carried out permutation testing and was originally proposed for EEG analysis by Blair and Karniski in 1993 (65). It has subsequently been used in several other studies although with slight variation each time (47, 48, 66–68). For example, either the data was only controlled for multiple comparisons over the time or channel domain (by taking the maximum initial-statistical value from each time sample or channel over time, and building several empirical null distributions). Or only a small portion of the collected data underwent the permutation procedure.

Since the critical value is calculated from the data, the critical value will be higher when there are high values in the original data since randomisations close to the original labels will also result in higher values. For a single channel permutation test this has no special consequences, however, for multiple comparisons, it leads to a test which is less capable of detecting lower, yet still interesting, secondary activations. This is a particular problem for EEG because signals tend to be stronger at the beginning of the ERP and weaker at later time points. Or spatially, dipoles near the cortex will produce a strong focal activation near the dipole, and a broad weak activation of opposite polarity on the other side of the scalp. Thus, although taking the maximum statistic to form the null distribution guarantees strict control over the false alarm rate, these tests tend to only find intense, focal differences in the ERP, and are generally found to lack overall sensitivity (69)

A proposed solution to ignoring the spatial and temporal information in the data is that one should first test each point separately for significance and then perform a second analysis looking for a minimum set of adjacent channels and/or time points which are also significant. Yet, this double testing of the data is both inefficient and statistically questionable. More importantly however is how one to determine just how many time points or channels should be next to one another, e.g. how does one decide that a 18 consecutive samples is just random activation, but 20 samples is a significant finding (33, 70–72)?

1.6.4 Cluster Size

Cluster size tests were introduced in order to include the important spatial information of the signal into a single test for significance (72). While the maximum-statistic approach calculates the significance threshold directly from the initial-statistics,

clustering involves two thresholding steps. Data clustering in the entire dataset can be readily detected by first setting a threshold on the initial-statistics and measuring the sizes of connected channels and samples that are above this threshold. For example, if two neighbouring channels are above threshold, each for two consecutive samples, we have a cluster size of four. For each permutated dataset the size of the largest cluster is taken in order to build the null distribution. The second *statistical* threshold is calculated from the empirical distribution of maximal cluster sizes in order to determine the minimal cluster size in the original dataset for significance. Important to note here is that regardless of the cluster-forming threshold, the permutation method ensures that the control for multiple compares remains exact. Clearly taking into account neighbouring information adds a very relevant aspect of the signal to the analysis and has been demonstrated to be generally more sensitive than the maximum-statistic approach (74, 75).

In using the cluster size to build the null distribution we are able to detect weaker, but more broadly distributed signals in both time and space and in this way is preferable to the maximum-statistic approach. However, because only the size of the cluster is measured, information about the channel's intensity is lost. Moreover, since a single statistic is now representative of a whole cluster, we can only draw conclusions about the cluster as a whole and cannot make direct inferences about the local maxima; thus inference is lost for specific channel-sample pairs. It should be noted that this is a rather severe limitation for fMRI data which has high spatial definition, but because EEG data has comparably poor spatial resolution, the main limitation is the loss of specificity in the time domain.

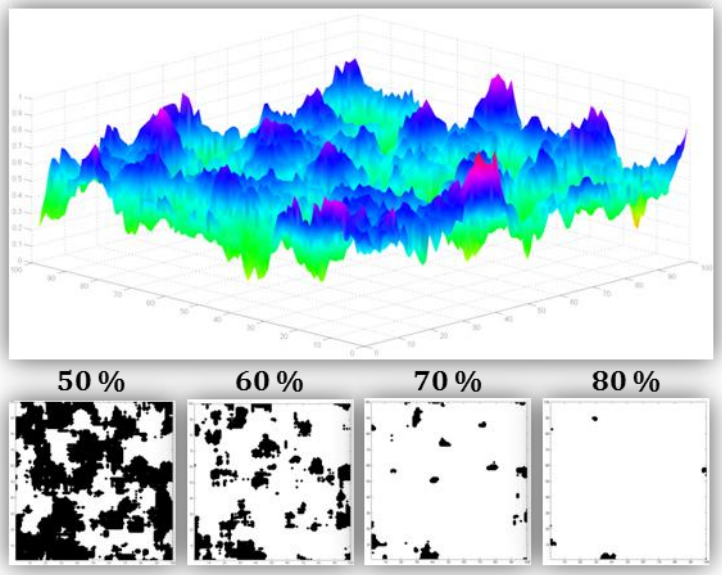
1.6.5 Cluster Mass

The cluster mass approach attempts to improve on the cluster size approach by calculating clusters over a certain threshold and then, for each cluster, calculating the sum of values over the threshold (74–76). Here, not only the size of the cluster is important, but also the actual intensity of values in the cluster. Thus, smaller-sized clusters but with stronger activation can be compared to larger, weaker signals. For example, a cluster of two channels where each channel is 1.5 over the threshold (cluster mass = 1.5+1.5), is equivalent to a cluster of three channels where each channel is only 1 over the threshold (cluster mass = 1+1+1). Although cluster mass statistics seem to solve the cluster-size issue of being exclusively sensitive to distributed signals over time and space, we are still losing information about the local maxima in the clusters.

1.6.6 The Threshold Problem

The larger concern for either the cluster size or cluster mass is the arbitrary selection of the initial cluster forming threshold value. In order to maintain statistical validity the only requirement is that the threshold be set a-priori to data analysis. However, the choice of threshold has dramatic consequences for the shape of the results as shown in Figure 3. In extreme cases, a low statistic will result in a single large cluster spanning the entire dataset, whereas an extremely high threshold may result in no clusters being found at all (77). For example, using a cluster mass technique Maris (75), found a large single significant cluster which spanned from 500-1500ms, which may indicate a threshold which was set too low. Although such a result is both statistical valid and sensitive; the extent of the single cluster makes the result nearly uninterpretable.

Figure 3 - Illustration of the threshold problem. Setting an arbitrary cut-offs of 50-80% has a dramatic effect on the thresholds found, and hence the appearance of the results.



In between these extremes lies a range of values that may result in useful and interpretable results (although it remains case that the thresholds in the low range will still be preferentially sensitive to large distributed results while higher thresholds would still miss larger but weaker clusters of differences). However, one cannot unequivocally determine which exact threshold will give useful results for a particular dataset, and so any cluster method remains likely to introduce user biases.

Rather than using direct t-value cut-offs, some have used uncorrected p-values from the known t-distribution. Although this may give slightly more stable results over different datasets (78–80), this represents inconsistent theory in that such a t-value cut-off

are calculated based on parametric assumptions which we have already argued are generally not met in the data. Moreover, even this threshold is likely to be arbitrarily selected since there is no theoretical justification as to why a p-value cut-off of 0.05 would produce the most optimal set of clustering in the dataset.

1.7 Chapter Conclusion

There is an abundance of available methods used to analyse ERP data[*]. What is still rather unclear is why the superior methods of SPM or the permutation techniques haven't completed taken over the field as they have with MRI. It may well be the product of 3 distinct factors. Firstly, EEG's long and slow development over the past 100 years has meant that analysis tools have always been a step behind technological possibilities. For MRI, the complexity of the data structure was immediately apparent and solutions had to be proposed while the technology was still being implemented. Secondly, recording EEG and creating ERPs is relatively simple and is often conducted by individuals with less technical experience. In contrast, the acquisition of MRI data generally requires a large team of people, many of which come from natural science or engineering backgrounds. Many researchers may opt for simpler techniques because more complex analytical techniques can be considerably more difficult to use, and the results more complex to understand. Lastly, conventional analysis is highly flexible and open to bias. This leads to an increased number of significant findings, which may or may not really exist. However, with the current state of research grants for funding, academic security and position, and scientific

[*] Methods described are by no means exhaustive. E.g. see appendix 6.4 for source analysis

pressure, methods with increased sensitivity may be too appealing to forgo.

Each method described has its own advantages and disadvantages. While some methods are clearly more acceptable than others, they all have some crucial faults which have held them back from becoming the mainstream analysis technique. In learning from the approaches several aspects for a candidate analysis method should be clear to be deemed optimal:

- All of the recorded data should be taken into account
    - But strict control of the multiple comparisons problem is essential
- The dependent measure used for analysis should be directly related to the data measured
    - That is, the magnitude and neighbourhood of the ERP
- Non-parametric statistics should be used to avoid the unattainable assumptions involved in parametric statistics
    - And also intuitively control for multiple comparisons
- Spatial information should be taken into account to increase sensitive to the common signals in EEG datasets
    - But arbitrary thresholding is a major issue
- The analysis should be simple to perform
    - No technical knowledge should be necessary in order to obtain an accurate result
- The output of the analysis should be easy to understand and be directly related to the original data
    - Researchers should be able to explore the results and share interpretations without having to go into detailed explanation of how they were made

Chapter 2 - TFCE

2.1 General principles

When we look at two ERPs, we have a general intuition about which differences we can trust to be really different and ones that are likely to be sporadic and random. This intuition is generally based on the magnitude of the difference, and whether the points' neighbours show a similar pattern of differences. In this way, we are likely to trust very large differences even when neighbouring points are quite different; but we would also believe smaller differences as long as there was a large neighbourhood of points that showed the same pattern. Methods which only take one of these aspects into account will not only lack in sensitivity, but usually require extra qualification of their findings to compensate for the missing aspect. Clearly, our intuition alone cannot be the basis of inference statistics. However, we can introduce common mathematical notation to describe that logic. Thus, we can denote the magnitude of the differences between two ERPs as 'h' (for height of the point of a waveform), and the size of the neighbourhood surrounding that point as 'e' (for extent of the cluster size). Following our intuition, the *real* difference between those two ERPs at a given channel and time point 'i' is some combination of our newly defined parameters 'e' and 'h'. Since we are likely to give different levels of importance to each of our two parameters, we can give each a specific weight to each factor denoted as capital 'E' and 'H' respectively.

The issues we have discussed with previous methods were faced in parallel in the field of MRI, and this mathematical justification of the intuitive difference approach is precisely what has recently been proposed in the MRI field by Smith and Nichols

(81). The goal is to enhance the initial statistic (e.g. the t-values), using both the intensity of the data point and information from neighbouring points. Following this, the enhanced statistic is further analysed by applying a maximum-statistic permutation method to control for multiple comparisons (see section 1.6.3 (82). The general idea is to enhance the value of weaker signals but which lie in large clusters to a level comparable with strong signals in smaller clusters. In doing so, the goal is simultaneously to suppress random noise that may have similar intensity as a true signal, but lack spatial-continuity. Importantly, because signals are enhanced individually and then analysed, the data will retain its local maxima and minima. Thus unlike all the clustering methods, all the information about peaks and troughs in the data is kept.

On a theoretical basis, this is done by calculating the supporting area under each data point. The supporting area is defined as the area underneath the point in the curve until its local minima. It is important to note that this is not a direct calculation of area under the curve but a point by point calculation akin to the calculation of the supporting structure for each point of a bridge. This value is then multiplied by the actual statistical intensity of the data point, defined by its height (h).

From a computational standpoint, this calculation is accomplished by applying a sufficiently large number of thresholds[*] to the dataset of initial statistics which then approximate to the integral of the supporting area. The thresholding procedure is applied in evenly spaced steps (s), between statistical values of 0

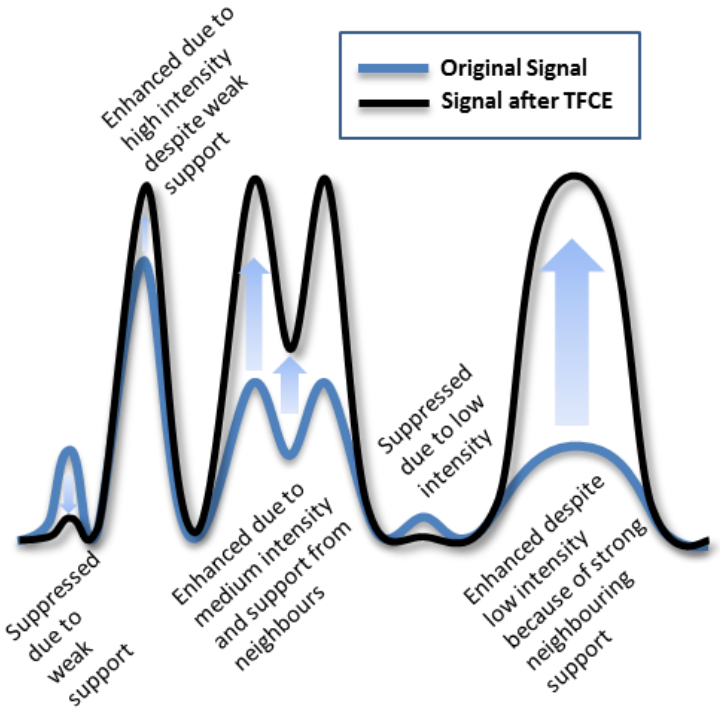
[*] The algorithm designed here uses 50 unique thresholds. This is more than sufficient to approximate to the actual supporting area. This number could be decreased if computational power is at a premium with the only effect of losing some accuracy in the final TFCE values.

and the maximum found in that dataset. For each threshold, neighbouring channels and time points are searched for above-threshold values. The cluster extent, e (raised to some power E), is multiplied by the threshold value, h (raised to some power H). Finally, the values for each threshold are summed up to form the new value for each data point (in that case of fMRI, each voxel, and in our case, each channel-sample pair). When we combine all the information into a single mathematical equation we obtain:

$$TFCE_i = \sum_{h=0,sh,2sh\ldots h_{max}} e_i^E \times h_i^H$$

This equation can be seen as the mathematical realisation of our intuitive notion of trustworthy differences. Moreover it can be seen a generalization of the previously described approaches. For example, the parameter settings of E=0, and H=0 will reproduce the maximum-statistic approach. Or, if the theoretical underpinnings of the cluster-mass technique seem sensible, but the thresholding problem is too large to overcome, the weighting parameters E=1, H=0, are the generalization of this idea without the single threshold requirement (81). If the equation is applied to each channel and time point of our initial-statistic we can obtain a fair assessment which includes the information about statistic intensity and its neighbourhood. In the end, different types of signals are enhanced or suppressed such that, despite the varied nature of the signal, we can compare their values numerically. The general effect of the algorithm is shown in Figure 4 below.

Figure 4 – Ideal outcome of the TFCE algorithm on different signal types. Data points with large supporting clusters will have comparable values to those with high peak intensities. This idea not only makes direct comparison between signal types possible, but also reintroduces the maximum statistic approach as a viable option to make inferences about the data.



2.2 The program

The algorithms used for calculation were mostly programmed in MATLAB® (The MathWorks Inc., Natick, MA). MATLAB was used as the programming environment because of its easy and effective handling of large datasets and the multitude of tools already designed. However, MATLAB actually tends to be

rather slow for repeated processing of individual variables, and thus the more basic programming language 'C' was used for the actual TFCE calculation and other clustering algorithms. The same algorithm written in 'C' may be orders of magnitude faster in comparison to even the most optimised scripts in MATLAB. The scripts written in 'C' are then interfaced with the existing scripts in MATLAB using so called 'mex-files'. These files must then be compiled for the specific computer configuration (e.g. 32 or 64bit computing). Section 6.2 of the appendix introduces pseudo-code of the computational algorithms which accomplish the analysis.

2.2.1 Inputs

There are only two inputs that are required for the program to perform its calculations. The first input is the actual data itself. Currently this needs to be organised into a single Matlab file for each experimental condition or group. This file should contain three dimensions; the first is a column-wise list of the independent observations in the study. For single-participant datasets, this would be a list of each trial, whereas for multi-subject studies this would be a list of the participants. Along the second dimension (columns) are the channels of the ERP; the order of which needs to be the same over all trials and participants. The third dimension of the data is therefore the time points collected which will depend on the sampling rate of the final ERP. Since there can be no missing data in the data, the number of channels and time points must be the same for each participant. Although this shape of the data may seem rather specific, most ERP data has this channel by time, or time by channel structure already. Including all the participants and possibly performing a quick reshaping of the dimensions is usually all that is required to take an output from any ERP creating software and analyse it using the method described here.

The second required input is the information about the locations of the channels. At the moment this should take the same format as the popular (and open-source), analysis toolbox EEGLAB (83) uses which is a structure file in MATLAB containing the channel label (e.g. 'FPz'), and sufficient information to localise the channel in 3D space. This may be in Cartesian coordinates where the position is specified by its X, Y, and Z points; or in spherical coordinates where two angles and the distance from the origin is provided. This information can either be systematically measured for each participant using infra-red tracking tools, or default coordinate files can be used. A further option, if the participant's MRI is available, is to fit the standard electrode model onto the participants structural MRI. The correct structure can still be easily created even if EEGLAB is not available.

Several further inputs can also be directly specified as options but can usually be left as default values or changed later if required. The default number of permutations is set to 2500. If computational power is an issue then this could be specified to a lower value. On the other hand, if the user desires an increase in the 'exactness' of the result, this could also be increased (see section 6.1 in the appendix for a discussion). Furthermore, the user can theoretically also specify the value of the E and H weighting parameters but this is not recommended or advertised and is included only for testing purposes (see section 2.5 on the optimal settings). The name for the results file can also be specified directly although the results are saved with a unique tag in either case. Moreover, the sampling rate and any baseline period can, and should be specified if an accurate time reference is desired. The latter two values can also be set later while viewing the results as they are not used in the actual analysis of the data.

2.2.2 Result viewer

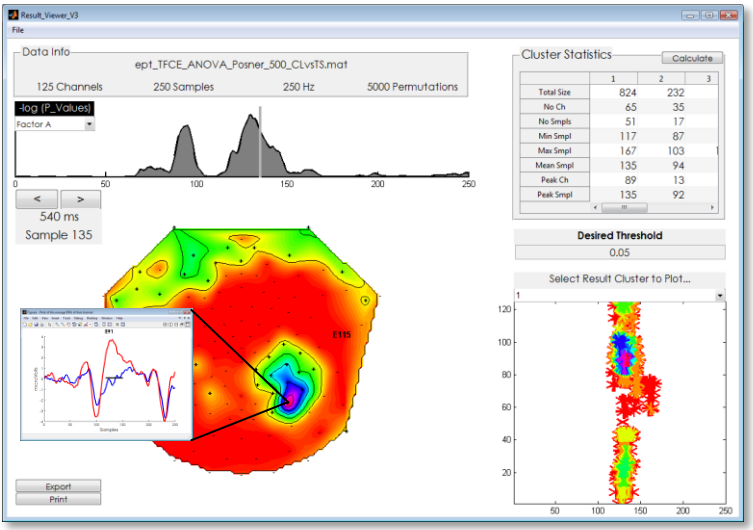
The results of the analysis are automatically stored in a file that can be readily accessed in MATLAB. The file contains several variables related to the analysis such as the original data, number of permutations used and the channel neighbours. The two most important of these variables are the calculated TFCE values of the observed data, and the calculated p-values of the data. Both of these are viewable as a channel by sample matrix where users can directly read off the p-value for a channel or time point of particular interest. Yet, with many channels and time points it is difficult to get an overview of the entire data in this way. For this reason a user-friendly viewing programme was created so that non-technical users of the analysis are nonetheless able to explore the results in an intuitive way.

The overall layout of the 'Result Viewer' can be seen in Figure 5. Upon starting the program the user can load any TFCE results file previously made. Upon loading, an area-graph is displayed which shows the sum of the negative log of each channels p-value over time. The calculation essentially gives smaller p-values an increasing number, such that peaks in the timeline correspond to the samples with the lowest p-values over all channels i.e. highest statistical strength. By default the plot shown is for the first factor found in the analysis. For simple analyses which only have one factor this menu is hidden. For more complex designs, factors and interaction effects are available in the drop-down menu.

The user can then proceed to click on the graph in order to show the topoplot of that sample. The topoplot includes either a '-' or a '+' indicating the channel position and whether this channel is significant or not at the specified significance threshold (default

setting at 0.05). The topoplot uses a triangulation algorithm to interpolate the values between channels fast and efficiently*. Left clicking over any channel reveals the channel's label. Right clicking over a channel opens a new window which shows the original ERPs for the selected factor, with an asterisk over the zero microvolt line representing the significant samples.

Figure 5 - Program used to view TFCE analysis results. See text for overview of features.



Finally, a table can be calculated which looks for connected significant channels and time points for a given significance cut-off. Several statistics are then given for each significant cluster found such as the total size, as well as its peak channel and sample. Another plot can then be created by selecting one of the clusters in

* Algorithm works up to 100 times faster than EEGLAB's topoplot function which is important when desiring to search through topoplots across samples quickly. The export button on the lower left export the topography in EEGLAB classical format which may be preferred in publications.

the drop-down menu. This plot visualised the entire cluster over all channels and time samples. The result-clusters make for a convenient way to report the scale of the analysis findings. It should be made clear that these cluster statistics in no way affect the statistical process and are only there to give another visualisation/summary of the results found by the TFCE process. Moreover, the data represented in these samples can be readily exported in order to create the visualisations in any another program (e.g. Microsoft Excel).

It is the fact of the TFCE method that the results are given in the same structure as the inputs which make the 'Result Viewer' so flexible. This flexibility and the ability for a user to easily explore different aspects of the results that makes the analysis accessible and interpretable. Any possibility to present the results that were open prior to the analysis are still open after the analysis but with the added benefit of being able to qualify them with statistical certainty[*].

2.3 Dipole simulation and sources overview

In order to properly evaluate a method, one should already know what the answer *should* be, and then see how close a particular method gets to that goal. To that effect, six varying signals were simulated using Patrick Berg's Dipole Simulator program (freely available at www.besa.de). Sources were created to give a wide range of intensities and cluster sizes in both the temporal and spatial domain. The simulated scalp potentials were taken from 129 electrodes in a geodesic array.

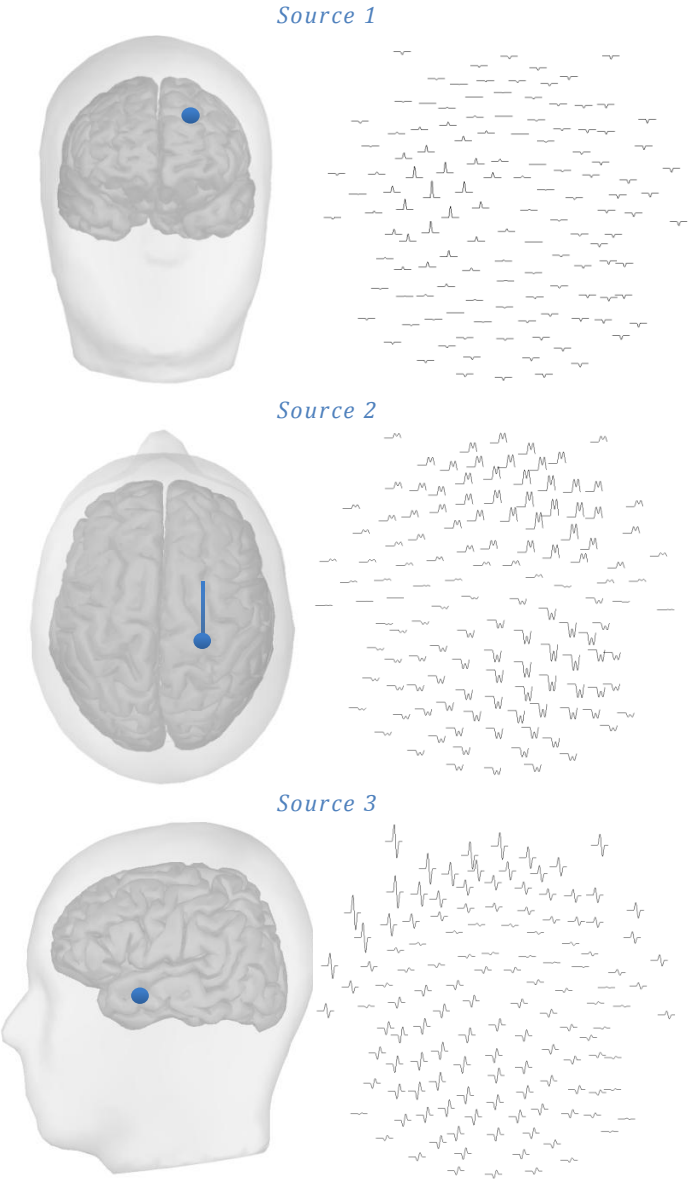
[*] For example, one could still perform source reconstruction (see section 6.4 or microstate analysis in order to visualise the data and aid interpretation of the results.

Noise datasets were also created using the Dipole Simulator which used 200 randomly located and randomly oriented dipoles. This method of noise generation leads to the special properties that the noise is correlated in time and topography with a frequency spectrum resembling of standard EEG with increased power in the alpha band. In order to assess the variability of the method, 10 complete datasets were created for each signal. A single dataset consisted of 36 signal plus noise trials.

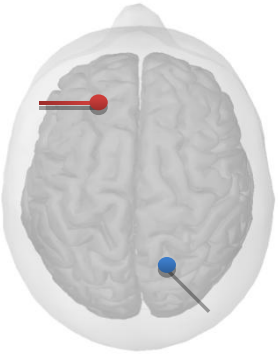
Furthermore, the effect of three different signal-to-noise ratios (SNR of 1, 2, and 5) were examined. The datasets' SNRs were controlled by normalising both the simulated signal and noise data and then adjusting the level of noise in a single trial. Since the averaging process in a dataset will approximately reduce the SNR by the square of the number of trials, the noise in a single trial was set at $\sqrt{36}$ times the desired SNR. Due to the fact that the normalisation occurs for the maximum peak of the original signal, the SNR value specified really only applies to the most intense peak in the signal. Therefore, the actual SNR for most of the data points is actually much lower.

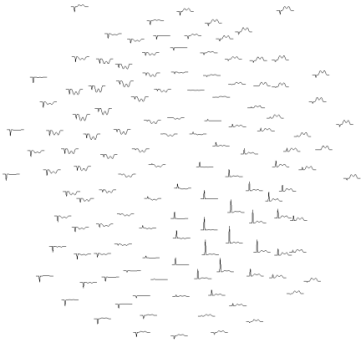
The first source represents a single dipole with focal activation. Source two has two fairly narrow peaks with a broad base and has a large distribution over the scalp. The third source is a single dipole with focal positive and negative deflections and a large distribution over the scalp. The fourth source contains two dipoles active at different times. The first of which has a short focal activation while the second's activation is identical to source two. The fifth source is similar to the fourth except the activity stems from just a single dipole at a different location. The final, sixth source represents the most complex pattern of activity. It consists of three separate dipoles in the frontal cortex, on the corpus-callosum, and parietal cortex. For this source, three types of

Figure 6 - Overview of Sources Used for Method Analysis

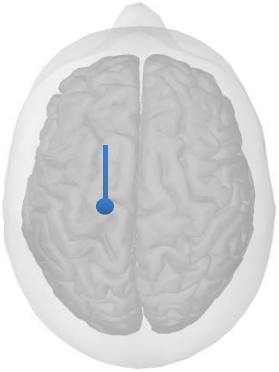


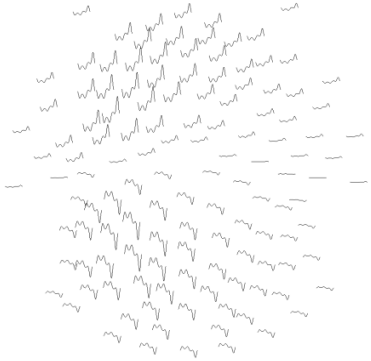
Source 4



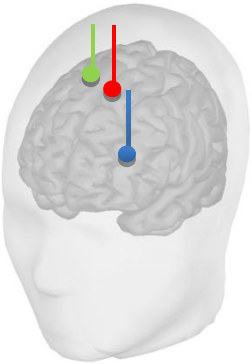


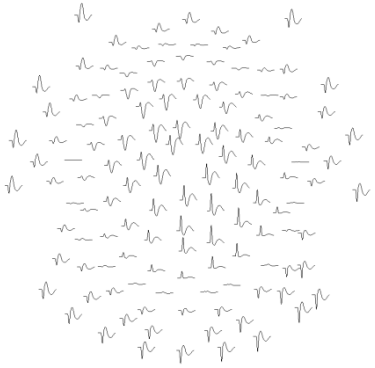
Source 5





Source 6





signals are portrayed: a focal but intense signal; a large but weak signal; and a medium signal in both aspects. The dipole locations and scalp topographies are further demonstrated in Figure 6.

2.4 Signal Detection Theory

Since the original is known, we are able to use tools from signal detection theory to make a formal assessment of the effectiveness of any method or parameters. In its simplest form there are four possible outcomes for any detection method (see Figure 7). On the one hand when a signal is present, the method may confirm this (correct hit, or true positive), or fail to find it (false negative). On the other hand, when a signal is absent, the method may confirm this (a correct rejection, or true negative), or claim there is a signal present (false positive). The proportion of each of these values can be combined in different ways to give a single statistic which defines the overall performance of a method. However, the case here is not a simple and straightforward case of signal detection and several issues need to be considered before we can define how to optimally define a good performance.

Figure 7 - Contingency tables outline the basic principles of signal detection theory.

		Ground Truth		
		yes	no	
Test Outcome	yes	True Positive (TP)	False Positive (FP)	**Precision** TP/(TP+FP)
	no	False Negative (FN)	True Negative (TN)	**NPV** TN/(TN+FN)
		Sensitivity TP/(TP+FN)	**Specificity** TN/(FP+TN)	

2.4.1 Definition of a true signal

In the strictest sense a true signal is detectable for any channel or time point where the activity is none-zero. However, with the signals having been generated by a true dipole source, virtually every channel over the dipole's active period has some activity over zero*. For example, in the dipole's active period of source 1, there are 564 data points above 0, but of these 43.6% are under 0.05 and 60.2% are under 0.1 (reminder: maximum signal strength has value of '1' since the sources are normalised). Considering that the SNR is defined by the maximum signal strength, it seems unreasonable to expect any method to find these very small signals; even in the lowest noise environments. If these smaller activations are considered part of the true signal, each method would suffer greatly in sensitivity and the differences between approaches diluted.

In the original TFCE paper for fMRI, Smith and Nichols recognised this issue and circumvented it by defining a true signal as those above 0.4 (a normalised value), and set the rest to zero. Although this would certainly address part of the problem just described, there are two severe problems with this approach. Firstly, this is a fairly drastic approach to simulated signals considering that, at least for the example of source 1 here this would mean eliminating 91.8% of the simulated signal. Secondly, setting a high truth cut-off will specifically eliminate the lower, larger clusters in the data and give preference to higher intensity peaks. This would in turn lead to a method that appears to function better for those signal types; perhaps acceptable for fMRI datasets but specifically unsuitable given the nature of normal EEG datasets.

* If a channel crosses a contour line perpendicular to the dipole, the activity will truly be zero, but this is relatively rare

Essentially we have a similar problem to the threshold problem (1.6.6 in cluster statistics and thus one we are attempting to solve with TFCE. Clearly we cannot rely on defining a true signal as anything above zero, or methods will appear rather insensitive. However, there is no clear cut-off above zero that does not come with other trade-offs. After some experimentation with truth-defining thresholds, the arbitrary cut-off point was set at 12.5%; this provided a reasonable balance between keeping most of the original signal but limited the bias of forming large clusters. That is, $1/8^{th}$ of the maximal signal in the source was defined as a 'real' signal and any channel-sample pair activity that was under this threshold was ignored. Assessment was also made for truth thresholds of 25%, 37.5% and 50% but most methods were able to detect signal under these thresholds and accurate discoveries were subsequently counted as false positives (when in fact they were *false* false-positives).

2.4.2 Binary versus continuous classifiers

Not only is the 'ground' truth actually much more appropriately defined on a continuous scale, our method does not simply output binary values of signal present or absent. Rather, each data point is assigned a corresponding p-value, which represents the degree to which we should trust the signal to really be different. Fortunately, signal detection theories have developed its more commonly used methods for these cases of continuous classifiers. Namely, the receiver-operating-characteristic (ROC) methodology allows us to compare tests for both their sensitivity and specificity by calculating and subsequently plotting one against the other (actually 1 minus the specificity), over a range of scores. The most common way of summarising this curve is to take the area under it (AUC).

Smith and Nichols (81) favoured a modified ROC approach in their simulations, but with some modifications. Rather than calculating the false positive rate (1 minus the specificity), from the simulated dataset itself, it was calculated from noise-only data. This gave them the advantage of not having to re-classify the ground truth after spatial smoothing since voxels near the true signal would have appeared significant. This effect is not of principle concern here, since we are conducting inferential tests which explicitly control the false positive rate[*]. Moreover, we are not smoothing the data prior to analysis and so any 'leaking' of the true signal which may occur due to the clustering methods, TFCE included, is of primary interest.

There are however two issues using the ROC methodology that are, at least in the case of our own data, too serious to use these methods of assessment. The first is that we are not dealing with a balanced dataset of signals. That is, we have far more true negative signals than true positive signals. This biases any method which tends to preferentially find negative results. For example, if our data contains 90% negative signals, then we would obtain an accuracy rating of 90% simply by calling everything we find negative. Moreover, this creates a situation in which the potential for false positives is much greater then false negatives. A quick look back at Figure 7 shows that most of the basic summarising measures are sensitive to this bias (especially specificity). Given a large bias, it becomes almost irrelevant how many false positives are found in the data, specificity is bound to remain high. Thus the value will immediately tend towards the ideal value of 1, and in turn, so will the ROC. Although methods could still be differentiated on this basis, it is difficult to reliably see differences in AUC which

[*] Smith and Nichols directly analysed the calculated TFCE values, without permutation or any other form of inferential testing.

are so close to the ideal. Without direct access to Smith and Nichol's dataset, we cannot confirm they were susceptible to imbalanced datasets, but given that for all comparisons except for the lowest SNR, most methods had an AUC indistinguishable from the ideal value, it can be assumed that they were.

The second point relates to the fact that in many cases the ROC curve is used to visualise the optimal test cut-off point to define a criteria whereby we determine if the test was positive or negative. This is usually done by finding the result threshold which lies closest to the top left corner of the ROC curve; as this would find the optimal sensitivity and specificity. However, because we are using inferential statistics (p-values), as our test measure, we already have a threshold criteria determined by the scientific community of *p* less than 0.05. Since values above this threshold are of no interest to us[*], we should create ROC curves only based on test values below 0.05. However, in any given test there may only be a few values (or none at all), which we can plot that are under 0.05 and the resulting ROC curve would have a rather poor sampling rate and be highly variable.

Taking all these issues account we propose three endpoints which, although none are ideal measures for reasons that will be discussed, in combination they present a good overview of performance. The first measure is sensitivity or recall at the p-value threshold of 0.05. Essentially this will indicate how much of the true signal was recovered in the test. This is chosen because we do already have some control over the false positive rate since we have

[*] It is in fact likely that the optimal combination of sensitivity and specificity lies above the p-value threshold of 0.05, but it is irrelevant in the scientific field if a certain method may perform best if we were to consider a p-value cut-off of 0.8 for example.

chosen the conservative norm p-value. The second measure chosen was precision, otherwise known as *positive predictive value,* also at the p-value threshold of 0.05. This will inform us about what percentage of the positive test results are actually true signals. Both measures have become popular alternatives to the more traditional ROC statistics because they are not as affected by imbalanced datasets (84–86). These two measures have an inherent trade-off and should always be examined together. For example, a test that always indicates a signal present will have perfect sensitivity, but at the cost of having lost all precision.

The final measure is special combination of the values from the contingency table known as Matthews Correlation Coefficient (MCC; (87, 88)), and is defined as:

$$\boldsymbol{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

As with the latter two measures of performance, the MCC was calculated for the threshold criteria of p = 0.05. As can be seen in the equation above, the MCC uses all four terms from the contingency table and gives an overall performance score, but is seen as more stable than other such measures for imbalanced datasets. The measure has been derived from theories of correlation testing such as the commonly used Pearson's correlation coefficient and essentially represents the correlation between the predicted values in the table and the observed values. The MCC outcome is between -1 and 1 where a value of 1 indicates perfect performance, 0 is a performance equivalent to random allocation of test results, and -1 is a perfect negative correlation. Moreover, it has been shown that the MCC can be directly converted into the Chi-square outcome measure χ^2 by squaring the MCC and multiplying it by the total number of observations.

2.4.3 Limitations of signal detection measures

We have argued in favour of our three selected measures, recall (sensitivity), precision and MCC as optimal methods given the available choices. However, for all three measures, using only the binary test classification that comes with using a single threshold criteria comes with a significant overall drawback that we do not know how these methods perform under this threshold. For example, a method which assigns a significant p-value of just below 0.05 to all the data is regarded as identical to one that assigns varying p-values based on the signal effect size. Thus, the chosen measures of assessment cannot be directly translated to compare methods if a researcher was more interested in significance values of under 0.01 for instance.

There is another very relevant limitation, not only for the measures chosen, but for all assessment measures discussed. Since they all rely on absolute counts of true positives from anywhere in the data, all methods of assessment will be more biased towards clustering methods. This is because there are more counts of true positives in signals with an extended range of time or channel points. Given that the same dipole source can produce extended or focal signals by even small changes in orientation or depth, it cannot be argued that the inherit bias towards cluster methods in signal detection parameters is a fair one. For example, orienting a near surface dipole perpendicular to the scalp may result in a strong focal signal over say 3 channels. Changing the orientation to parallel to the scalp will result in a weaker signal but over say 15 channels. When only counting true positive signals, the first orientation would result in only 3 true data points, whereas the second in 15 true data points. If these two signals were present in the same dataset, the pure cluster method would find 15 of the 18 total true positive signals whereas the intensity sensitive method would only find the 3

of the 18. However in truth, each method has only been able to detect the consequence of a single dipole out of the two. Thus, there exist easily identifiable cases were this bias might overwhelm performance measures and the limitation should be kept in mind when reviewing the resulting tests. In other words, we must keep in mind that methods which favour the detection of larger clusters may not actually perform as well as the assessment measures may indicate.

2.5 Optimal Values of E and H

Although we have removed the need for a single arbitrary cluster-forming threshold, we have introduced two new weighing parameters, E and H. This may seem like a step in the wrong direction but as Smith and Nichols demonstrated these values could be set to non-arbitrary defaults which could be determined empirically and have solid roots in several statistical theories (81). For a logical perspective, if the value of H was set to 0 each successive cluster measured would carry the same weight in the final sum, but since the clusters formed at the lower thresholds would naturally be much larger they would dominate the final sum. More intuitive is that clusters formed at increasingly higher thresholds should be given increasing importance, such that H should be larger than 0. Furthermore, when considering the initial-statistic as t-values, we know that increases in t-value do not follow a linear increase in proportion to their importance. That is, a t-value of 4 carries more than just double the significance compared to a value of 2. Therefore, H should be larger than 1 to reflect this non-linearity.

When considering an appropriate value for E, remember that when using low cluster-forming thresholds we are likely to find

very large clusters which span over a great deal of channels and time points. These large clusters at low thresholds hold little importance in practice in determining signal differences; thus, we should want to limit their importance by choosing a value of E less than 1. From a theoretical perspective Appendix B and C of the original paper for fMRI present a detailed discussion on the theoretical optimal values of E and H (81). Briefly, transforming well known weighted p-norm functions, and estimating negative log p-values using random Gaussian fields suggest the values of 2/3 for the weighting of cluster extent, and 2 for the weighting of the intensity parameter.

Here we attempt to empirically derive at the optimal weighing parameters for general EEG data using simulated data, detection theory and 5 different setting for both E (0, 1/3, 2/3, 1, 2) and H (0, 1/2, 1, 2, 4), giving 25 different possible combinations. All of the six sources generated were tested, for each of the 10 datasets for each source. This results in 250 TFCE tests for each source and a total of 1500 tests run for this assessment. Inference was made on the cumulative rank of each parameter setting for each source. In other words, the results of recall, precision, and MCC were ranked from worst to best performance and then summed up for each source. 2500 permutations were conducted in each test.

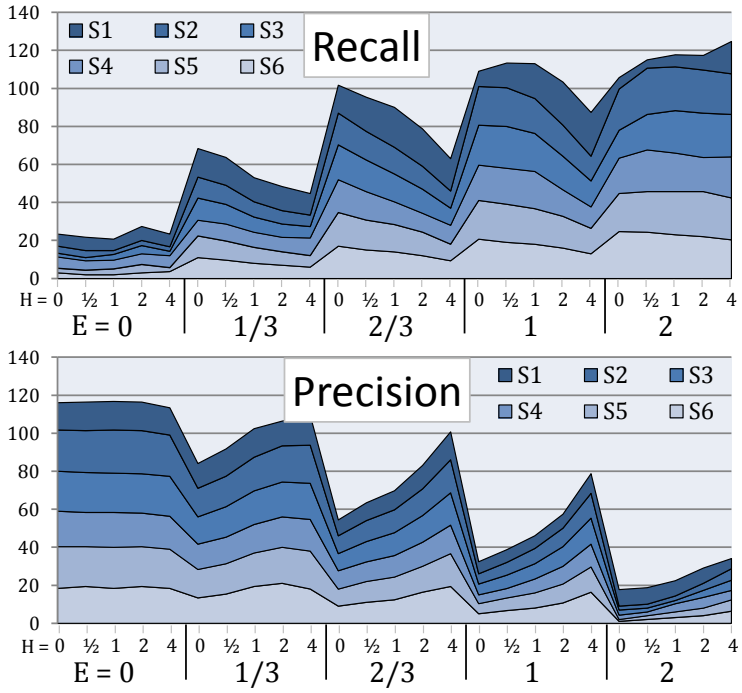
2.5.1 Simulation Results

Despite the variation in signal sources and SNR, there was a clearly discernable pattern of performance measures over the different parameters. Several general findings are summarised below:

- Precision and recall show remarkably similar patterns in parameter ranking regardless of the SNR in the data.

- Recall progressively increases when the cluster extent is given a higher weighting by increasing the E parameter.
- Recall is reduced when higher intensities are given more priority by increasing the H parameter. The exception being for cases where E = 2 for higher SNRs.
- Precision generally increases with when the H parameter increases. Except in the cases where E = 0.
- Higher values of E reduce the precision of the tests.
- For increases in SNR, the optimal settings, as defined by the MCC, generally decrease in the E parameter and increase in the H parameter.

Figure 8 – Average recall and precision for each source. The y-axis represents the cumulative rank of the 25 parameter combinations tested. Ranks are averaged over each SNR value of 1, 2, and 5.



Due to the nearly identical pattern of recall and precision over all SNR values, Figure 8 shows the mean ranking of recall and precision measurements over the three SNR levels tested. The MCC values on the other hand displayed different trends dependent on the data's SNR level and thus Figure 9 shows the overall performance measure MCC over each SNR. More specifically when looking at the actual values calculated, for the lowest SNR of 1, average recall is fairly poor across all parameters, ranging from just 0.4% to 20.7% of the data discovered (mean 8.5% ± 7.2). Precision on the other hand is kept to near ceiling levels for most sources, with the average ranging from 99.4% to 81.0% (mean = 94.1% ± 6.5). Thus, all the parameter settings remain essentially quite conservative, sacrificing the number of significant points found but at least being fairly sure about those. Looking at overall performance, the MCC value has its maximum for both average MCC (aMCC), value and rank at the setting E = 1, and H = 1 (aMCC 0.170 ± 0c.102). Although slightly lower rank, parameters E = 2, H = 4, achieves essentially the same average MCC value but with reduced variability over the datasets (aMCC 0.170 ± 0.094).

For the mid-range SNR of 2, the average recall improves considerably (mean 33.4% ± 17.2), but also with a now considerable range of values from 7.7% to 50.4%. Despite the large increase in sensitivity, precision remains fairly high (mean 87.3% ± 11.8), although values range from 99.7% to as low as 66.5%. The highest MCC value belongs to the parameter settings of E = 1, H = 2 (aMCC 0.488 ± 0.077). Not far behind in average rank, and with less variability across the datasets is the parameters E = 2/3 and H = 1 (aMCC 0.475 ± 0.067).

With the SNR at 5, and thus with the most easily identifiable signal, recall ranges from 46.8% to 93.4% with a relatively high mean of 76.7% (± 16.1), of the true signal detected. As with the

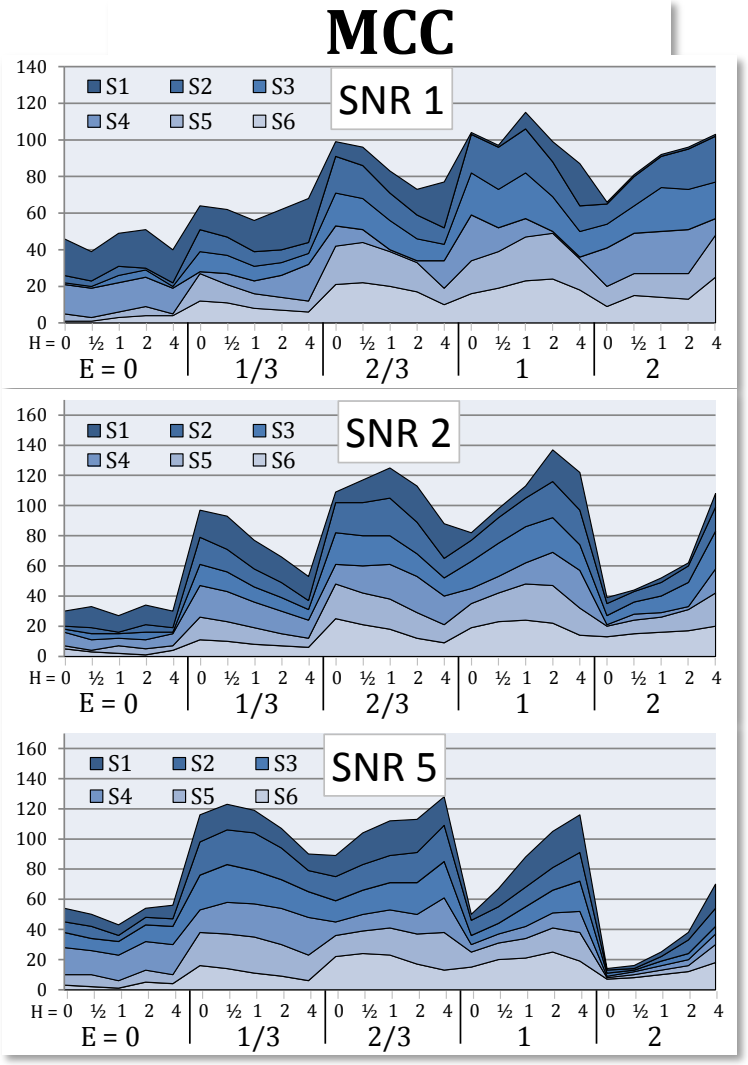
other SNR values, precision is maximal for parameters where E = 0 at 98.5% and lowest for high values of E at 58.5% (mean 79.0% ± 13.9). Initially it may seem odd that precision decreases with increasing SNR, but in percentage terms this is not totally unexpected given the large increase in sensitivity and absolute number of true positives. Here, the most optimal method over all the sources is the parameters E = 2/3 and H = 4 (aMCC 0.768 ± 0.030). However, several other methods perform almost equally well; E = 1/3, H = 0, 1/2, and 1 (aMCC 0.764 ± 0.030); as well as E = 2/3, H = 2 (aMCC 0.762, ± 0.031); and E = 1, H = 4 (aMCC 0.762 ± 0.029).

Looking across all SNR values the top five parameter settings in in order of aMCC and rank were E = 1, H = 2 (aMCC 0.462 ± 0.076), E = 2/3, H = 1 (aMCC 0.457 ± 0.067), E = 1, H = 4 (aMCC 0.454 ± 0.057), E = 2/3, H = 0.5 (aMCC ± 0.080), and finally the theoretically determined value of E = 2/3, H = 2 (aMCC 0.448 ± 0.057).

2.5.2 The ideal weighting for E and H

All parameters tested perform fairly well, with the clear exception of E =0 settings (ignoring neighbourhood information), and E = 2, H = 0 (high cluster weighting with no regard for signal intensity). Thus it seems that as long as there is some weighting given to both the E and H parameters, tests will perform fairly well. On the grounds of this empirical test alone, the optimal settings appear to be E = 1, H = 2 since it scored consistently high for all SNR values, and was the best overall. Moreover, this setting has a decent sensitivity to the various signal sources (mean Recall 49.1% ± 8.4) while maintaining high precision of the results (mean Precision 83.9% ± 8.9). However, there are several reasons why we remain more drawn to the theoretically derived values of E = 2/3, H = 2. As discussed in 2.4.2 the performance assessments methods are biased towards

Figure 9 – Cumulative rank of Matthews Correlation Coefficient (MCC) over SNR values of 1, 2, and 5, for each source. Y-axis is the sum of the parameter ranking out of the 25 parameters of E and H tested. Note that the ideal parameters show a decrease in E and an increase in H as the SNR improves.



clustering methods since there would be more true positives found. Therefore a slight reduction in the E parameter or increase in the H parameter is closer to ideal given the known limitations in our assessment method. Additionally, the assessment measures essentially balance the importance of recall and precision, whereas in the research field, we tend to favour more conservative approaches if all else is even. As Figure 8 demonstrated, a decrease in E or increase in H is associated with higher precision values.

These arguments could just as well speak for parameters E = 1, H = 4. Judging from the pattern of the results, it certainly seems as though we can obtain similar performance by keeping a certain ratio between the two parameters rather than giving some definitive values to each setting. For example, settings of E = 2, and H = 8, although not explicitly tested, could be expected to perform just as well. But since this 'fixed-ratio' hypothesis would require further testing, and possibly a mathematical proof to be conclusive, we should stick to the ratio that works well and also currently has the most theoretical backing.

The parameters E = 2/3 and H = 2 should remain as set defaults that require a solid theoretical and empirical reasoning in order to be changed. Firstly, because TFCE is, in terms of overall performance, relatively stable to changes in reasonable parameter settings anyways. Secondly, because the signal type is generally unknown prior to running the experiment and so changes would likely come after first analysis and thus bias the result. And finally because our aim is to eliminate these user biases and create a standardised methodology whereby results can be directly compared.

A persuasive argument could be made to 'tweak' the parameter settings if the SNR was known. SNR cannot strictly be

known from real datasets since it would imply the precise true signal in the data was already known, and hence no need for further statistics. However, we can obtain an estimate of the SNR in the averaging process by additional calculating two averages taken from the odd and even numbered trials in the data. Assuming a single trial is just a linear mixture of the real signal and random noise, and assuming the true signal in the data remains constant throughout all the trials, subtracting odd and even averages will eliminate the true signal and leave an estimate of the noise with zero-mean with some noise variation. The real averaged signal on the other hand should substantially reduce the noise and enhance the true signal present in the data. The SNR of the dataset can then be estimated by dividing the signal variance by the noise variance. Using this estimate, one might be theoretically justified in changing the parameters of E and H to more optimal parameters. Namely a higher ratio of E to H values for low SNR and vice versa for higher SNR estimates. However, it is important to realise that these estimates may only be reliable for a sufficient number of trials where SNR is expected to be fairly high; and thus would ultimately lead to using the default values in any case.

2.6 The Effect of Filtering

Frequency filters of some kind are almost always used in EEG pre-processing and will inevitably change the shape of the data. In order to determine what effect this may have for the TFCE algorithm, raw data from source 6 at SNR 2, underwent 5 different filters. Source 6 was chosen because it contained 3 separate signals with different wavelengths. Filters of 10, 20, 30, 40 and 50 hertz, low-pass, zero-phase, 16th order, Buttersworth IIR filters were designed in MATLAB and applied to the data. Ten datasets were

formed and analysed as in the previous section but using TFCE with only the set parameters of E = 2/3 and H = 2.

Table 2 summarises the finding of the performance assessment. Almost every filter used, with the exception of the 10Hz filter, improves the MCC value. The filters do so essentially by increasing signal sensitivity, while maintaining a high precision of results. The filters are able to improve sensitivity by reducing the power of high frequency noise in the data, and thus improving the SNR. The 10Hz filter fails to improve the SNR, not because it fails to reduce the high frequency noise, but because one of the source signals has a wavelength of higher than 10Hz. Thus, the power of the true signal is also reduced and comparable levels of SNR to the no filter situation are achieved. As long as the filter does not impede in the same frequency range, the clustering extent of the data will increase and higher TFCE scores will result. From a theoretical perspective TFCE should be stable to different filtering approaches. Different filters will, by design, either reduce the intensity of a signal, but increase smoothness and so cluster extent, or leave the signal intensity intact but at the cost of jagged peaks and valleys which ultimately reduce the cluster extent. Filters thus have inherent trade-offs but ones which TFCE is sensitive to and thus the approach gives comparable results across filters. Even if the results are achieved through different weighting of intensity and cluster size.

The single clear drawback of any filtering is however noticeable in the consequent variability of the created datasets. This is particularly noticeable for sensitivity where the standard deviation more than doubles in comparison to the no filtering situation. Thus filtering brings the potential gain of increased signal sensitivity, but by no means guarantees it. However, the noise in these datasets was modelled with a maximal frequency of 125Hz,

and with a decreasing slope in frequency power. With real EEG recording, there are multiple sources of noise at potentially higher frequencies, and with increased power in those higher frequency bands. Therefore, conservative filtering techniques should always be used when possible as real noise often spans the entire frequency spectrum.

Table 2 - Assessment results for 10, 20, 30, 40, and 50Hz low-pass filters on datasets of source 6 at SNR of 2. The shaded bars in the background provide a visual companion to compare values (e.g. Precision values do not actually differ substantially).

	Recall	*std*	Prec.	*std*	MCC	*std*
10Hz	0.228	*0.100*	0.944	*0.042*	0.309	*0.069*
20Hz	0.252	*0.092*	0.977	*0.025*	0.352	*0.071*
30Hz	0.238	*0.093*	0.980	*0.021*	0.341	*0.069*
40Hz	0.231	*0.090*	0.977	*0.022*	0.333	*0.066*
50Hz	0.220	*0.090*	0.978	*0.024*	0.324	*0.066*
No Filter	0.206	*0.042*	0.981	*0.018*	0.318	*0.031*

2.7 Chapter Summary

So far we have demonstrated the theoretical superiority of the TFCE method in several ways. Firstly, it was shown how the TFCE equation can be seen as the mathematical formulation of the idea that both the size of the cluster and the individual magnitude of the point are important factors. From this it was shown that other methods such as the maximum-statistic and the cluster approaches could be generalised in the TFCE equation by using different weighting parameters. In doing so however, it could be directly seen that those weightings were not optimal from a logical perspective. Theory would suggest a value of E = 1 or less so that lower threshold clusters did not dominate the results. In turn H

should be higher than 1 since the intensity values are generally based on non-linear statistics and higher threshold clusters should be given a higher priority than those at a lower level.

Dipole simulations were created for different signal types and noise levels in order to empirically determine optimal values for E and H. Generally, it was shown that increasing the weighting of cluster extent subsequently increases signal sensitivity. Whereas increasing the importance of higher level clusters increases the precision of the methods. Performance measures indicated that the optimal parameter settings over all sources and noise levels were E = 1, and H = 2. The theoretically derived parameter settings of E = 2/3 and H = 2 also performed well, and arguments were made for those values to remain the default settings. It was also shown the method is relatively stable to different filtering parameters, and results could be improved with moderate, commonly-used filter settings.

Furthermore, a brief overview of the analysis program was given; as well as a look at the guided-user-interface implemented to allow easy exploration the procedure's results. The result viewer highlights how the method creates interpretable results which can be directly reflected back to the original ERP waveforms. More than this, the resulting topography and significance strength over time can be quickly examined to give an overview of the entirety of the experimental effects.

Chapter 3 - Direct Comparison to Other Methods

Chapter 1 demonstrated the limitations of each of the currently applied summary statistics in permutation. In particular, the cluster mass, cluster size, and the maximum-statistic approach. Chapter 2 then introduced the TFCE approach and demonstrated how it overcomes these issues on a theoretical basis. Yet, even a conceptually optimal method will not be accepted unless it is also shown to be useful on a practical basis. To this end, the TFCE method was compared against many others using both simulated sources and several examples of real data. By comparing the methods ability to detect significant differences arising from a simulated source, we are able to give precise statements about how the method compares to others in terms of its sensitivity and specificity. Using real data from external sources, we can confirm the methods reliability in real-world applications. Despite the obvious benefits of direct comparison between methods, and the abundance of available statistical techniques for EEG, there have been relatively few publications that have attempted the feat. Two recent reviews are discussed below, followed by our own comparisons using the optimal TFCE settings discovered in the previous chapter.

3.1 Previous Work on Method Comparisons

In the first of two larger comparison studies, Lage-Castellanos and co-workers compared four different methods of false discovery control to the analysis of ERP data (89). The maximum-statistic approach under permutation was compared

against the simple uni-variate test statistics (the uncorrected t-distribution), as well as two proposed methods of global and local false discovery rate methods. In essence, these tests are all a form of intensity based statistics since they are designed to give a feasible significance threshold to the t-values calculated. Using single and multiple channel simulations of ERPs, they found that the maximum-statistic approach with permutations performed too conservatively compared to the other methods.

There are three main issues with this work that bring the practical relevance into question. The first is that their simulations of datasets do not actually resemble that of real EEG. The two sinusoidal waveforms they generated are typical of ERPs but the waves remain the exact same over the selected channels and show no activity over the others. Thus there are no localisable peaks and valleys in their simulations over topography. This simulation method certainly makes the results easier to compare, but they actually may tell us little about the methods real potential. The second issue is that they used only the false-discovery rate and power measurements to assess the methods without an overall score which balances the two findings. Although it may be easy to create a mental scale of the measures in clear cases, the trade-off between the two factors cannot always be demonstrated fairly without a further measurement. Lastly, the comparisons are only made for intensity based measures of which none take the spatial and correlated structure of ERP data into account. As already discussed in section 1.6 these types of methods are already quite well known for their lack of sensitivity and over-conservativeness and pointing out the rather subtle differences between these sub-optimal methods does not substantially further our knowledge on the subject.

In a more recent article by Groppe et al, several multiple correction methods were tested using simulated ERP data (90). This study differed to the previous one discussed in that the cluster mass statistics proposed by Maris and Oostenveld were also included (75); and different performance measures were used. Here no one particular method is proposed as optimal, but each method is discussed in its own right. In general the simulation results were consistent with theory in that the maximum-statistic approach proved to have excellent control of the false-positive rate at the cost of sensitivity, especially to distributed signals. The cluster-based permutation tests only provided weak control over the number of false positives, and were especially powerful at detecting weak broadly clustered signals. For their conclusion, an interesting table is given as a guide to which method should be chosen given certain a priori hypotheses and assumptions about the data.

The simulation comparison in this thesis has several advantages to those previously published. The first is that several different sources are used with realistic dipole sources of both the signals and the corresponding noise. Three different SNR values are used to examine how each method behaves in these conditions. Various cluster-forming-thresholds (CLFs), are used to examine the influence of this arbitrary but crucial choice for cluster-based methods. Moreover, not only are the methods compared for their sensitivity and power, but an overall assessment method is used in order to make decisions about which method is indeed the most optimal.

3.2 Simulated Source Comparison

3.2.1 Source data

Using the same simulated sources already described in 2.3 the currently used summary statistics in permutation were compared against the TFCE approach. Two parameter settings were taken for the TFCE weighting factors of E and H. The first, TFCE-A, where E = 2/3, and H = 2, corresponds to the theoretically derived optimal settings; and TFCE-B, where E = 1, and H = 2, corresponded to the optimal settings empirically when tested in 2.5

Ten datasets were created for each source, with 36 signal and baseline trials for each dataset. Noise level was adjusted for each dataset to create three distinguishable SNR levels of 1, 2 and 5. Then each method analysed all the datasets using 2500 permutations for each dataset. For cluster approaches, cluster forming thresholds (CFTs), were set at t-values of 1, 2, 3, and 4 to see how each threshold affected the results over the various sources and SNR levels.

The same performance measures were used as described in section 0to determine how close each method came to the known true signal: recall of the signal (also known as sensitivity); overall precision in the results (percentage of recovered signal that actually corresponds to true signal; and finally, MCC, a correlation measure of expected versus discovered signal which uses each value in the contingency table and thus gives an overall measure of performance.

3.2.2 The results

- Despite the variation in source, each method is surprisingly consistent in its ranking for precision and recall. Source specificity is essentially only seen in the MCC values.
- The trade-off between recall and precision is clear for all methods.
- Cluster mass and size perform very similarly. Their results seem to further converge when SNR improves.
- Both the cluster mass and size methods perform optimally at a cluster-forming-threshold of 2 for this data.
- When increasing the CFT for both clustering methods, the sensitivity to the signal decreases but precision improves.
- The maximum statistic approach is overly conservative. It is almost always the least sensitive of all methods, but subsequently has the highest precision.
- Cluster methods with a high CFT essentially behave like maximum-statistic methods.

Recall measures vary significantly over sources and SNR (e.g. cluster mass with CFT 2, does not find any significant data for the first source, but achieves a 60% recall for source 2 at SNR 1). Mean recall values overall were 7.6% (± 6.3), 27.3% (± 7.3%), and 64.3% (± 5.8%) for SNR 1, 2 and 5 respectively. Cluster mass with CFT 1 achieves the maximum recall levels over all sources for both SNR 1 and 2 (21.5% ± 14.9%; 51.5% ± 14.6% respectively), but with considerable variability between datasets and sources. For the highest SNR tests, TFCE-B achieves the maximum signal sensitivity over all sources (91.3% ± 4.3%).

Figure 10 - Average recall and precision for each source. The y-axis represents the cumulative rank of the 11 methods tested; TFCE (A, E=2/3, H=2; B, E=1, H=2; Cluster Mass (CM) and Cluster Size (CS) at cluster forming thresholds of t=1, 2, 3 and 4; and the maximum-statistic (MaxT) approach. Ranks are averaged over each SNR value of 1, 2, and 5.

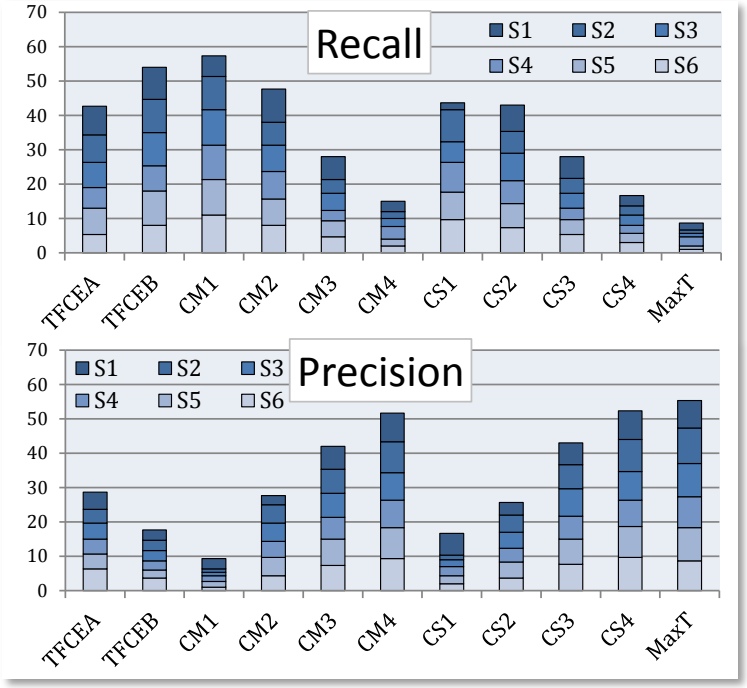
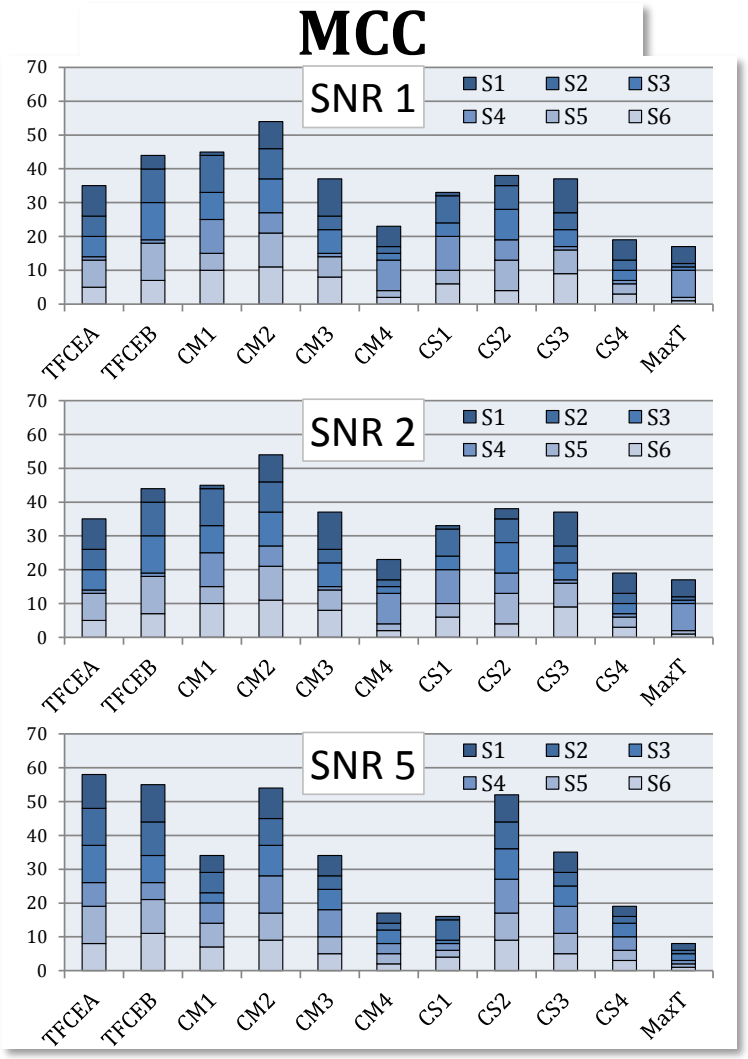


Figure 11 - Cumulative rank of Matthews Correlation Coefficient (MCC) over SNR values of 1, 2, and 5, for each source. Y-axis is the sum of the parameter ranking out of the 11 methods tested; TFCE (A, E=2/3, H=2; B, E=1, H=2; Cluster Mass (CM) and Cluster Size (CS) at cluster forming thresholds of t=1, 2, 3 and 4; and the maximum-statistic (MaxT) approach.



Precision remained at fairly high levels for most methods over sources and SNR levels; a direct consequence of setting our performance measures at a p-value of 0.05. Precision tended to actually decrease when improving SNR. For SNR of 1 mean precision over all sources was 95.9% (± 5.8%), for SNR of 2 the value dropped to 91.6% (7.4%) then further to 89.7% (± 3.4%) for SNR of 5. This is likely to do with two factors. The first drop is likely caused by the fact that with increased sensitivity, there will be a higher chance of obtaining false positives from a percentage point of view. The decrease for SNR of 5 on the other hand is more likely to do with the fact that several methods are able to detect signal under the 12.5% cut-off for a true signal, which would result in a *false* false-positive (see 2.4.1 for the definition of a true signal). This explanation is supported by the fact that the standard deviation does indeed decrease with increased SNR as expected.

Due to the inherent trade-off between recall and precision, the MCC result is the determining factor to assess overall performance. For the low SNR of 1, the MCC ranges from 0.031 (MaxT), to a high of 0.165 (TFCE-B) with a mean of 0.113 (± 0.076). For the slightly higher SNR of 2, values range from 0.170 (MaxT) to 0.488 (TFCE-B) with a mean of 0.339 (± 0.063). Here the TFCE-A method was a close second with an MCC of 0.455. For the highest SNR of 5, values ranged from 0.442 (Cluster Size with CFT of 1), to 0.762 (TFCE-A) with a mean of 0.635 (± 0.049).

As shown in Table 3; when averaging over all sources and signals the top three methods found were the TFCE-B (E=1, H=2), the cluster mass with CFT of 2, and the TFCE-A (E=2/3, H=2), method in descending order. However, the reverse order is true when looking at the variability of each of these methods, with TFCE-A method being the most consistent of the three over all sources and SNR levels.

Table 3 – Mean value and standard deviation (std) for performance measures recall, precision, and Mathews correlation coefficient (MCC) for each of 11 methods tested (see Figure 8 – Average recall and precision for each source. The y-axis represents the cumulative rank of the 25 parameter combinations tested. Ranks are averaged over each SNR value of 1, 2, and 5. above for details). Each cell corresponds to the mean over all sources and signal to noise ratio. The top three values are highlighted in bold.

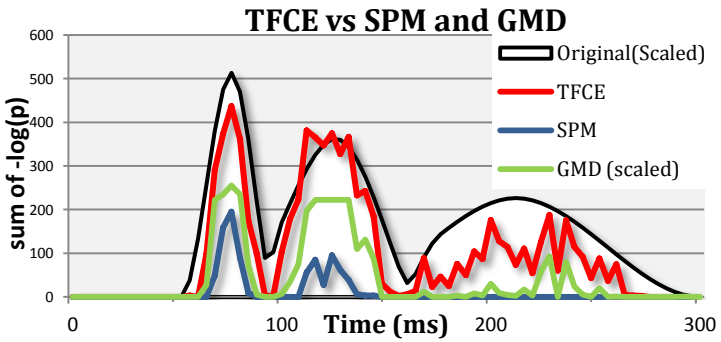
Method	Recall		Precision		MCC	
	Mean	*std*	Mean	*std*	Mean	*std*
TFCEA	**0.416**	*0.058*	0.906	*0.051*	**0.448**	*0.057*
TFCEB	**0.491**	*0.084*	0.839	*0.089*	**0.462**	*0.076*
CM1	**0.535**	*0.119*	0.747	*0.137*	0.377	*0.103*
CM2	0.402	*0.059*	0.939	*0.062*	**0.457**	*0.065*
CM3	0.271	*0.036*	0.992	*0.011*	0.374	*0.046*
CM4	0.183	*0.023*	**0.998**	*0.005*	0.286	*0.034*
CS1	0.379	*0.192*	0.820	*0.157*	0.269	*0.130*
CS2	0.387	*0.067*	0.939	*0.068*	0.432	*0.076*
CS3	0.271	*0.036*	0.991	*0.013*	0.372	*0.045*
CS4	0.183	*0.023*	**0.998**	*0.004*	0.286	*0.033*
MaxT	0.126	*0.015*	**0.996**	*0.013*	0.220	*0.024*

3.3 Direct comparison to SPM and GMD

SPM is currently one of the only available parametric methods that is able to test for significance effects over all channels and time points (see 1.5). For this reason it is also crucial to know how this approach to multiple comparisons correction behaves for simulated sources, especially in comparison to the TFCE method presented. Furthermore, microstate analysis, and its commonly used statistical measure, global map dissimilarity (GMD) have often been used in publications (see section 1.2 In this brief analysis, significance was tested on a single dataset from source 6 at the

medium SNR level of 2 using permutation statistics on the GMD calculation as well as taking the p-values directly from the SPM analysis. This simulated dataset was chosen because it contained three separate dipole signals representing large focal signals, weak but largely distributed signals as well as a signal with medium sized and intensity. Furthermore, each signal dipole was located in a different region of the brain and thus this dataset represented the most varied and complex combination of sources available for comparison. For the SPM results, a peak-level family-wise error (FWE), correction, based on random field theory, was set at p = 0.05 and no cluster extent threshold was set.

Figure 12 – Comparison of the TFCE method, SPM, and GMD in the analysis of source 6 at an SNR of 2. The sum of negative log of significance values over all channels are presented so that a higher value relates to a higher total significance. In the case of GMD, values needed to be scaled up since GMD is a single measure over time. SPM significance values are taken from the family-wise error correction based on random field theory. The scaled global field power of the original simulated signal is also shown.



SPM found six significant voxel clusters of which only three were over 100 voxels in volume (FWE correction for clusters at 0.05 would have been a minimum size of 280). These three large clusters found, essentially represented the positive and negative deflections of the first early dipole signal, as well as the central-positive deflection of the second dipole. The smaller three clusters were

scattered representations of the boundary-negative cluster of the second source dipole. SPM found no significant voxels for the third simulated dipole which represented a weak, but largely distributed signal. In comparison to the known true signal, SPM found just 5.5% of the total true signal, however had a ceiling precision level of 1 (no false positives), and with that an overall MCC score of 0.193. Since the GMD is a reductive measure over all channels; no specific information on the sensitivity or specificity can be given that is comparable. For the GMD statistical comparison, only 11 of the 75 total samples were found to be show topographical difference from its random noise baseline condition. Of these 11 none were found in the later portion of the ERP where the weak but broadly distributed signal was generated. Thus the GMD measure appears to be insensitive to the broadly distributed weak signals.

The TFCE method using the theoretically derived values for E and H of 2/3 and 2 respectively, for the same dataset, had a recall of 28.0%, a high precision of 97.7%, for an overall MCC score of 0.444[*]. Importantly, TFCE found significant channel-sample pairs for all three of the simulated dipoles in the signal, including the later weak but broadly distributed signal SPM completely failed to detect. Moreover, the TFCE method not only showed a larger total number of significant channel-sample pairs but also a generally higher level of significance for true positives. This can be seen in Figure 12 which shows higher total significance (as the sum of the negative log of the p-values over channels), over each time point. The p-values obtained by the TFCE approach most closely resemble the shape of the global field power of the original simulated EEG signal. It should be noted that since it is not possible to extract

[*] In fact 18 of the 21 false positives found here were actually part of the true signal but under the 12.5% defined cut-off for truth. Thus precision and MCC scores are in reality substantially higher.

specific p-values for particular voxels in a result dataset, SPM p-values were estimated by using the results of decreasing FWE corrected results.

3.4 Discussion on simulation results

The results here demonstrate the practical power the TFCE approach has for various kinds of signals that may be present in an EEG dataset. The optimal parameter setting found in the previous chapter, TFCE-B was shown to be the overall optimal method over all sources and SNR values. The theoretically derived parameter settings, TFCE-A, was also shown to be a particularly powerful method, ranking third overall with a higher precision at the cost of sensitivity to the signals. This in spite of the assessment measures being biased towards clustering methods.

The cluster-based techniques performed generally well with a CFT of 2, with the cluster-mass technique coming in second place over all methods. Proponents of the technique however still face two major issues before the method could seriously be regarded as a generally accepted analysis. The first is to make an argument that a CFT setting of 2 was an optimal setting prior to conducting the analysis (see section 1.6.6). The second is that despite the method finding many true positives in the data it is unable to determine where the most significant peaks are in the data. This is because the cluster as a whole is given a p-value and not the individual data points. For example, for source six, where three dipoles create separate but slightly overlapping signals, the cluster mass approach finds only two large clusters of significant data spanning the entire time span corresponding to the positive and negative deflections on the scalp. Since the p-values for these clusters are 0.004 and 0.008 it is not possible to determine, on the

grounds of these results alone where the most significant points are in the data. Strictly speaking one would have to interpret these results as a single experimental effect spanning the length of the ERP. Therefore from a signal sensitivity perspective, the cluster-based techniques may perform extremely well, but they achieve this at the expensive cost of validity and interpretability.

Although the SPM and GMD analyses were only performed on a single dataset, the comparative results of the TFCE approach were far superior in sensitivity and MCC score to that of SPM. Given SPM's similar approach to the maximum statistic technique, it is likely that its performance substantially increases for higher levels of SNR. The GMD measure seemed to perform slightly better than SPM but since neither specific electrode configuration or statistical neighbourhood is taken into account, GMD still performed far worse than the TFCE measure for this dataset. Moreover, with the GMD measure, we have reduced all the channels to a single measure for analysis, and hence also for results and we cannot make inferential statements about an individual channel's contribution to the GMD parameter. This is a rather steep price to pay for data that is nevertheless ultimately less accurate than TFCE.

The fact that these comparisons were limited to a single dataset could be seen as a potential bias; however the dataset was selected based on the fact that it represented the most EEG-like and complex simulated signal. Furthermore, only a single analysis was conducted because of the increased dataset preparation time needed in SPM to obtain a result. Moreover, custom scripts needed to be written for SPM to obtain the FWE-corrected p-values for each channel-sample pair.

Thus far we have shown that the TFCE method is valid in its statistical framework; it is superior in its theory by being a

generalised expression of other methods, with parameter settings with solid theoretical and empirical backing; and now the theory has been shown to be optimal in its sensitivity and precision over a range of signal types and SNR values. However, this superiority has only been demonstrated in the controlled setting of simulated sources. It may be still be argued that the range of sources and SNR values were particularly tuned for TFCE, or that EEG data does not really look like the simulated signals. Moreover, one may be curious as to how the method deals with data in the frequency spectrum. To approach the issues, in the subsequent sections, real data is analysed which was obtained from external sources and further compared to previously introduced methods.

3.5 Real data from SPM

3.5.1 Data source

All three datasets explored in subsequent sections were obtained from free online databases. The first real dataset comes from SPM's tutorial on EEG analysis of single subject data and is described in detail in chapter 36 of the SPM manual (EEG mismatch negativity data). 128-channel EEG was measured from a single participant while they performed an auditory odd-ball paradigm with 480 standard tones (500Hz) and 120 rare tones (550Hz). Raw data was average referenced, down-sampled to 200Hz, and epochs were created from 100ms before event onset to 400ms post event (101 total samples). Trials with artefacts were rejected after a simple threshold detection algorithm set to 80 μV leaving 437 *standard* trials and 107 *rare* artefact-free trials to be compared.

Here, two analyses were carried out to compare several different methods using high and low signal to noise ratios (SNR) in the first dataset. For the high SNR version of the dataset, all

available trials were used. For the low SNR analysis, only the first 20% of the trials were used from both conditions leaving 88 *standard* and only 20 *rare* trials. It is relevant to examine different levels of SNR as this directly translates the sensitivity any method has. Moreover, if a method can be shown to work at low SNR then paradigms could be constructed with fewer trials being necessary in order to elicit the correct statistical result.

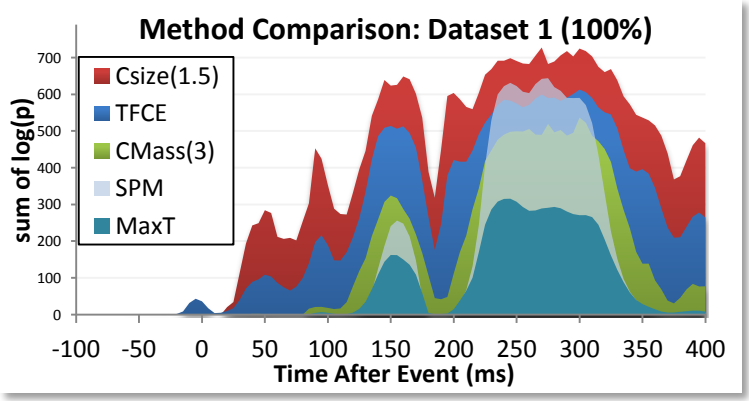
3.5.2 Results

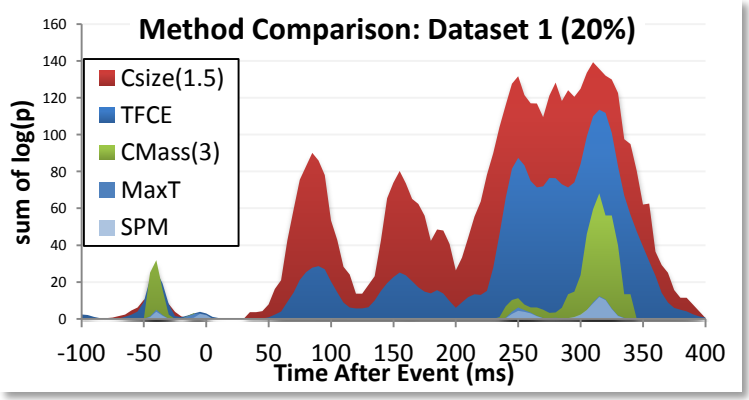
Figure 13 shows an overview of the main findings of a selected few methods. The original SPM analysis of 100% of the trials showed 6 significant clusters of data with the largest having their peaks around 300, 270 and 160ms. The highest significant voxel was found nearest to channel A3 at 300ms. The second largest cluster corresponded to the negative deflection of the first cluster. For the cluster around 160ms, the peak significant voxel was found nearest to channel C1. The clusters essentially form two larger effects of interested. A much smaller cluster of only 2 voxels was also found around 100ms.

The two smallest clusters found had sizes of 6 and 2 channel time-pairs which indicated that essentially any clustering in the data at that threshold is deemed significant. The permutation distribution for these approaches indicated that for most random relabelings no channel-time pair was over T = 4.5 and so the 95% cut-off was zero, meaning even a single channel-time pair would have been significant. This explains why the approaches demonstrated essentially identical results as the maximum statistic approach. The difference in significance power between the two approaches stems from the fact that the specific p-values for the cluster approach were the minimum possible value while the maximum statistic approach the p-values were not quite minimal. It

is worth repeating that although mean time points were given for clusters, there are no intensity peaks in p-value since these are the same for all data points within a cluster.

Figure 13 – Overview of selected methods in their analysis of 100% of the SPM mismatch negativity tutorial dataset (top), as well as when using only the first 20% of the trials from the same dataset (bottom). Graphs depict the sum of the negative log of the obtained p-values over all recorded channels such that a higher value is indicative of lower p-values or more channels with significant differences.





Finally the TFCE approach yielded 5 significant clusters. The largest cluster (2147 channel-time pairs) around 260ms was comparable in size to the low threshold approach. The second largest cluster (1180 channel-time pairs), had a mean time point of 196ms but upon further inspection there were clearly two clusters joined by a small, minimally significant (mean significance $p = 0.0115$), chain of channels at the time points between 190-220ms. The earlier peak in the cluster occurred at 165ms at channel A1 (Cz), whereas the later peak was found on channel A13 (posterior left) at 255ms. A third large significant cluster (745 channel-time pairs), which essentially represented the opposite sign of the first largest cluster, peaked at 310ms. Interestingly, the TFCE approach also yielded two smaller significant clusters centred at 50ms and 75ms which no other method reported (although included in the low-CFT cluster approaches). Although inspection of the ERPs would suggest these are real differences, although without some ground-truth it cannot be known whether they are false positives or not.

In summary, the low threshold cluster size approaches essentially found all data points to be significant after 20ms. The higher threshold cluster approaches performed like a simpler maximum statistic approach since any channel-time pair over the threshold was found to be significant. The SPM, TFCE, and optimised cluster approach all performed similarly with the TFCE approach being equally sensitive to the main differences at 160ms and 300ms, while SPM and cluster methods showed preferentially sensitivity to the later component. Furthermore, the TFCE approach found earlier significant channels that no other approach found.

Figure 13 (bottom) gives an overview of the comparison when using only the first 20% of the data. SPM analysis showed a single small significant cluster (13 voxels) at 315ms. As with high SNR data the cluster size and cluster mass techniques performed similarly, finding essentially the same significant data but with

different p-values. Low threshold (T = 1.5) found a single large cluster (1241 channel-time pairs) from 40ms to 400ms, with the cluster mass approach finding a higher p-value for this same cluster. The medium threshold (T = 3.0), yielded a single cluster (85 channel-time pairs) at 320ms with both approaches having similar p-values. When using a high CFT (T = 4.5), only the cluster size approach found a single significant channel across two time points (315ms and 320ms). The maximum statistic method found no significant channel-time pairs for this reduced dataset.

The TFCE approach also yielded a single significant cluster, but of 257 channel-sample pairs involving 25 unique channels and ranging from 235ms to 365ms. Upon further inspection of the cluster, again, two separate peaks were clearly present. The earlier peak occurred at 260ms around electrode D20 (left temporal) while the later peak occurred around electrode A3 (left central-posterior) at 320ms.

3.6 Real Frequency Analysis from SPM

3.6.1 Data source

This dataset represents a single channel from a MEG experiment exploring the perception of faces to scrambled pictures and again comes from the SPM tutorial (Chapter 37 – Multimodal face-evoked responses). In this study 275 MEG channels were measured, epoched around the events from 200ms prior to 600ms post event (161 samples), then baseline corrected and down sampled to 200Hz as per instructions in the SPM manual. As in SPM, it is possible to compute and analyse all channels in the time-frequency bands but a single channel was taken in order to directly compare the results of the TFCE approach to the SPM analysis described in the SPM manual.

3.6.2 Results

Both the SPM and the TFCE found two similar significant clusters of data. For the lower frequency cluster, SPM found 79 frequency-time pairs with peaks at 5Hz and 185ms. The TFCE approach found all the same data points to be significant but included a further 19 frequency-time pairs, and had its peak at 5Hz and 190ms. For the slightly higher frequency cluster SPM found only 32 frequency-time pairs to be significant while the TFCE approach found 98 frequency-time pairs in the same region to be under the significance threshold. For this cluster, both methods had the same frequency and time peaks at 12Hz and 100ms.

3.7 Reanalysis of previously published group data

3.7.1 Data source

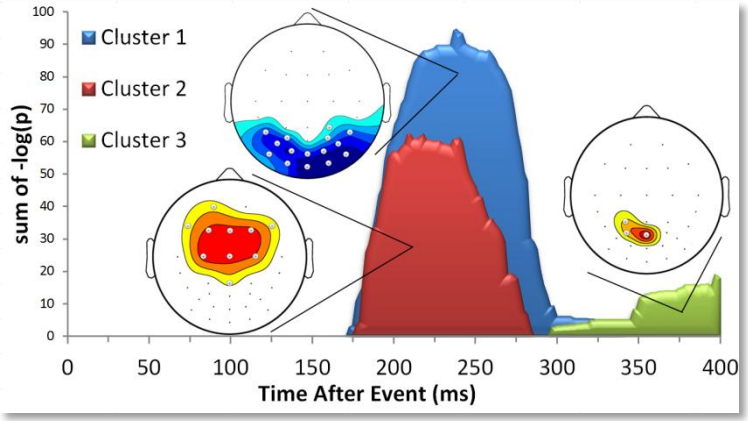
This dataset is taken from a study where 14 total participants performed a go/nogo animal categorisation task and demonstrates the approach for a paired group comparison. The dataset has been previously analysed using different methods (91, 92). In the task, participants were required to release a button whenever a briefly-presented picture contained an animal. EEG was recorded from 32 channels from the international 10-20 system, average referenced and down sampled to 500Hz. Each participant's ERP was created by averaging over animal and non-animal trials from 100ms before and 400ms after picture onset (250 samples).

3.7.2 Results

Analysis found 3 clusters of significant ($p<0.05$) data. The largest cluster (710 channel-time pairs) ranged from 172ms to 340ms. Its peak occurred at channel at 238ms and represented an

increased positive amplitude for non-target distracters primarily over posterior electrodes (T_{13}= 9.31, p= 0.0005). The second largest cluster (441 channel-time pairs), ranged from 176ms to 284ms. Its peak occurred at 208ms on channel F4 and roughly corresponded to the mirror, frontal activity of the first larger cluster (T_{13}= 6.71, p= 0.0015). The third, smaller cluster (106 channel-time pairs), was found between 296ms to 398ms. The peak here was found to be at 382ms at channel POz (T_{13}= 8.49, p= 0.0015). From its location on the scalp and longer latency this likely represents a secondary signal whereby non-target distracters had a lesser positive amplitude than targets. The earlier deflections at 98ms and 120ms, described as significant in the paper, were not found using the TFCE method with the smallest p-values at 98ms and 120 being 0.51 and 0.42 respectively. Figure 14 shows the three clusters found to be significant and the topography at the time points of the highest sum of the negative log p-values over the 32 channels measured.

Figure 14 – Summary of significant results of the TFCE reanalysis of group comparison. Significant clusters are calculated post-analysis by searching the p-value structure for connected channels or time points that are under the specified significance threshold (in this case 0.05). Three clusters of significant data were found with different peaks of effect.



3.8 Discussion on real data results

Results from all analyses essentially indicated that the TFCE approach, in terms of sensitivity and specificity, performed on par or better than other approaches for single-subject channel-time data, at different SNRs, as well as frequency-channel data and group comparisons.

For high SNR data all approaches yielded some significant results although several differences in result structure were immediately apparent. The maximum-statistic approach is the most basic of all methods since information about the relationship between channels in time and space is not taken into account, and it is therefore not surprising that this method was the least sensitive to differences in conditions. Setting a CFT to 4.5 was equivalent to the maximum statistic approach since most permutation datasets only found much lower t-values. Thus, the permutation distribution found that the minimum cluster size necessary for significance was 1 channel-sample pair. Thus, setting the CFT to an arbitrary 4.5, in this case, is analogous to arbitrarily setting the t-value significance threshold to 4.5, a procedure which clearly has no statistical validity. In fact, the 5% cut-off value of the permutation distribution for the maximum statistic was 4.489. Therefore, even though the algorithm by which the high-threshold cluster approach is calculated is statistically valid, the process is essentially equivalent to a non-valid one; and a CFT of 4.5 could be considered a lucky guess. The TFCE approach, because every data point is calculated, always has some maximum value for each permuted dataset, and as such will always have a distinct distribution with consistent peaks and valleys and will circumvent the floor effects described above.

On the other hand using a relatively low threshold to form clusters is extremely sensitive to neighbouring information and

even minimal consistency in the data values will result in significant data points. Considering the high correlation between neighbouring channels in EEG, especially with high density recordings, minimal consistency in their statistical values can almost be guaranteed. The finding of significant differences already after just 20ms post stimulus is also indicative of an increased false positive rate using such a low threshold. Perhaps most importantly is that the low threshold approach highlights one of the principle weaknesses of the cluster approach in that since every data point in a cluster is given the same p-value, there are no individual localisable peaks in the data. Thus, even for the high SNR data we would only be correct in concluding that the experimental manipulation had a significant effect after 20ms over most channels. Due to this weakness, it is not uncommon for location of the maximum t-values to be expressed as the local maxima in large clusters. However, even within a single cluster there could theoretically be a single high t-value surrounded by very low values while in another region in space or time there may be a larger group of only slightly lower values which would be more statistically relevant. Therefore, although the procedure is statistically valid, an arbitrarily low threshold is likely to be yield results which are overly general and thus uninterpretable. Once again, since the TFCE approach enhances every data point by its supporting clusters, local maxima are retained in larger clusters. Thus even in large areas of significant data, local peaks are identifiable on the basis of their precise p-values with the significant cluster. Furthermore initially smaller values, but with large supporting clusters, may be enhanced to values higher than a single intense channel. This allows for a direct comparison of the different values within and between clusters of significant channel-sample pairs.

For low SNR, the low threshold cluster size approaches seem to be the most sensitive as the significance strength (Figure

13) does indicate that more channels were found to be significant around the time points described for high SNR. However, only a single cluster was found and thus the same issues apply with interpretation of the results as described for the high SNR results. Moreover, because significant differences were found as early as 40ms, which likely reflect false-positives, there would be no statistical basis for deciding which of the channels could be interpreted as real experimental differences. In this case, neither the maximum statistic nor the higher cluster threshold approaches yielded any significant results since with a lower SNR, the t-statistic did not reach the required threshold. These issues from the cluster approaches raise two important points. Firstly, that the optimal threshold does not only depend on the type of signal to be detected, but, even if the type of signal to be detected is known, the SNR of the data will also play a role in determining the useful range of thresholding values. Secondly, although theoretically the cluster mass approach seems like a better alternative than simply taking cluster size, results from both approaches over multiple thresholds have often been essentially identical. This is due to two important steps in the process: firstly the initial threshold will yield the same clusters for both approaches; secondly, at least in this dataset there are essentially only quite small clusters (under 20 channel-time pairs), or quite large clusters (over 300 channel-time pairs), and so the critical cut-off essentially just eliminates the smaller clusters were both size and mass will be under the threshold. Therefore, whether the cluster mass approach is an improvement over the cluster size actually depends on the distribution of cluster sizes and the amount of mass left in the cluster once thresholded.

In the secondary analysis on SPM's tutorial data using frequency analysis the TFCE analysis showed increased sensitivity over the SPM analysis by finding a larger frequency and time range

to be significant. It maintained its specificity however, assuming the SPM analysis can be taken as the ground truth, in that the point of maximum significance was equal in both analyses. The analysis was carried out on a single channel in order to directly compare the results with SPM, however, it is possible, and we recommend, to carry out the analysis on all the channels, time points, and a wide frequency range in order to fully explore the data and see whether the effects found are similar or even more pronounced in neighbouring channels.

The group data produced similar results as were published in the original paper (92). However, in the published analysis, researchers grouped channels into frontal and occipital regions and corrected for multiple comparisons using an arbitrary value of 15 consecutive time points. Moreover, using the TFCE method, we are able to give channel and time peak information based on the location of the highest significance level between the two conditions, as opposed to the maximal difference in amplitude given in the original results. This is important as amplitude differences do not show how much variability the two conditions had, nor do they include information from neighbouring channels that may make a peak in a well supported cluster more significant than a higher isolated difference in a single channel. Furthermore, early differences at 100ms reported as significant in the paper were not found to even show a trend in our analysis. We think it is more likely that their significant finding there was likely to be a consequence of the multiple testing and insufficient correction since in the original data only a single channel (Oz), showed relatively high t-values between conditions in that early time range. However, the TFCE analysis did find a third significant time range in posterior channels later in the ERP which the original paper did not test for. Thus, using the TFCE method, much of the same conclusions could have been drawn but with more confidence in the

statistical results, more specificity in localising maximum effects in time and space, and without the risk of over-interpreting differences from a single channel.

Common to all approaches, except TFCE, is that potentially crucial information is being ignored in the data. Spatial information is neglected in the maximum statistic approach while for threshold-dependent clustering approaches, information about intensity and the other thresholds is lost. The TFCE approach is thus unique as information from all relevant thresholds are included, as well as the actual value of the channel measurement. As a consequence the p-values obtained do not need to be qualified by further measures and can be regarded as a direct measure of how confident we can be in the differences between signals. The key issue with the TFCE approach is not which information to include, but rather how each piece of information is to be optimally weighted. For these real datasets we have shown that the default values for our weighting parameters E and H provide good results for high and low SNR versions of the same dataset, as well as for frequency analysis and a group analysis. In showing its efficiency, any deviation from the default values would have to be strictly justified when using the method, unlike the choice for cluster-forming thresholds which cannot have an overall optimal value across various SNRs and signal types.

Lastly, with no further input requirements after specification of the ERP data and channel locations, the approach is automatic and generates a full set of results within a few minutes. Thus, exploration of the resulting structure can be done easily and intuitively without the explicit need to understand the details of the statistical process underlying those results.

Chapter 4 - Expansion to Complex Designs

The previous chapters focused on single comparisons between two groups/conditions. However, modern research designs rarely just focus on just one factor and will often want to explore the role multiple factors may have in a certain experimental setting. Although a researcher may be initially tempted to run multiple single comparisons for each experimental hypothesis, this will lead to a new multiple comparisons problem and more importantly, will not be able to show how certain factors may interact with one another; or be able to control for the variation of other factors when looking at main effects[*]. Thus, for the TFCE approach to become useful in everyday experimental settings, a method which can analyse multiple factors requires exploration. First the theoretical aspects are considered when randomising across several factors with varying underlining designs. Then a shortened description of a research project is given in which those theoretical concerns are applied.

4.1 Considerations for complex designs

For large datasets containing several correlated measurements, as is typical in current EEG datasets, it is nearly impossible for the assumptions of parametric statistics to even come close to being acceptable. As already discussed, these

[*] This may also be one of the reasons the conventional analysis has remained popular since many studies will run an ANOVA on their experimental factors and simply include a few channels and samples as additional factors in the general linear model.

assumptions should be questioned even when assessing the outcome of a simple t-test between two datasets. However, they become even further untenable once we introduce multiple factors into our statistical model. For this reason, inferential testing using the permutation method is also highly suitable in this area.

For more complex designs such as an experiment testing for differences between two groups for two conditions (i.e. a 2x2 mixed design), situation quickly arise where an exact permutation test may not be possible. Furthermore, it may not be clear just what the exchangeable units are in the design for specific hypothesis testing. For example, if the experimenter is interested in an exact test for a single factor, then only the units of that factor can be exchanged under permutation and the others must be left constant. In that case, each permutation for the second factor will result in identical values (since no randomisation is taking place) and thus inferences cannot be made about this secondary factor. In most cases however, a researcher conducts a more complex design in order to examine the main effects of the factors and any possible interactions between them. Here, exact tests are no longer possible since certain coefficients in the general linear model will always have to be estimated. However, one can still obtain very accurate estimations of each factor and interaction without the need for additional assumptions.

The suggestion by Manly (93) is to directly permute all raw observed values and then randomly reassign it to the groups. That is, if we have a simple 2x2 between-subjects design with 5 people per group. The procedure would then simply be to take out all 20 observed values, shuffle them, and then randomly reassign values to create identically shaped 2x2 cell blocks, yet with different participant data. For 2x2 repeated measures designs, the 4 condition blocks should be exchanged within the same participant.

This is because with repeated measures, the conditions for each participant are bound to correlate and as such cannot be regarded as independent observations and are not exchangeable across participants. The same logic is applied for the permutation of data in paired t-tests. For mixed designs the same reasoning is applied and raw data permutation is a two-stage process. First, data for the repeated measure is exchanged within the individual participants, and then that participants' data is exchanged across group labels. This results in full permutation whereby observations from a single participant remain together, albeit in a possibly different order, but their group labels may differ (94). In other words, the observations from any single participant will never end up as observations in two participants under randomisation.

Anderson and colleagues (95–98), have taken a more complex approach based on a different perspective on the null hypothesis. In their view, it is that the error terms in the complex model should be equal across groups or conditions. Thus, it is the error terms that are the exchangeable units. Generally, for any ANOVA test under the permutation approach, the exchangeable unit is the denominator term of the F-ratio of the test. Strictly following this general rule provides an exact test for any individual term in the model. However, the exactness of one factor would mean fixing the errors of another factor in the model and a fixed term under permutation will result in identical randomisations for that factor; thus excluding the possibility of finding differences for that factor. Following this logic it is clear why no exact test exists for an interaction term since that would imply fixing all main effects, and thus leaving no units which can be exchanged.

In order to test multiple factors and their interactions simultaneously approximate methods have to be used. Simulations have shown that the most powerful method for calculating

interaction terms is permutation of residuals which can control for the main effects of factors. That is, the full ANOVA model is calculated for the original data, and then for each individual observation, the residual is calculated by subtracting from its value the means of each factor and level, then adding the overall mean. This is essentially equivalent to finding out how much data remains unexplained from each observation if the original model is correct. The residuals are then permuted and an empirical null distribution is calculated. Although strictly approximate, this method of permutation of residuals is said to be 'asymptotically exact' since although the direct influence of the main effects are not kept constant, there variability is kept fixed by removing their mean values. Thus these tests are not approximate in the same way that parametric tests are said to be approximate.

However, permutation of residuals, by controlling certain main effects, does not then provide significance testing for them. Anderson (98), argues that testing for main effects only makes sense when the interaction is non-significant. There, since the interaction would have been shown to be non-significant, it's term can be removed from the model and exact main effects can be tested. This may be plausible for a single dependent variable, but for mass univariate statistics, as in our EEG data, interactions will not be significant for the entire dataset and it will always be necessary to see if for those channel-samples pairs factors' main effects are. Moreover, in mixed random factor designs, it may of interest whether main effects are significant, over and above any more specific interaction effects.

Therefore, for a more exploratory analysis, where both the significance of main effects and their interaction(s) are of interest, the only available solution is to run a permutation of the raw observations as discussed earlier. This procedure has been shown to

retain strong control of false positives over most data types and experimental designs, but can often lack power to detect real effects (94, 98). In this sense, we are edging on the side of being overly conservative using this permutation approach; which although not optimal is the more preferable error type to make in science. For experimental designs where a single factor is of primary interest, and other factors are introduced to control for some known influence, permutation of residuals may provide exact control of false positives and at the same time provide the most sensitivity to effects.

4.2 Posner Paradigm

One of the most popular tasks for examining the orienting of attention was designed by Posner and colleagues in the 1980s (99–102). In its most basic form, the task is essentially one of reaction time where a participant is asked to respond as quickly as possible to a target when appearing usually in either the left or right visual field. Several factors have been examined which has significant effects on participants reaction time. Two of the most investigated factors are the influence of a cue-event prior to the presentation of the target, and the time interval between the cue presentation and the subsequent target (stimulus-onset-asynchrony; SOA). As expected, reaction times tend to be faster for validly cued trials and longer SOAs. However, when a re-orienting event takes place between initial cue and target, an effect known as inhibition-of-return (IOR) is observed. In IOR validly cued trials tend to display a slower reaction time than other trial types for longer SOAs. This is thought to be caused by the bias of attention away from previously explored locations but is by no means the conclusive interpretation (103, 104). Furthermore, studies have found significant effects on reaction times when cue where either

explicitly or implicitly directing attention. Practically, this aspect has been manipulated in two ways: either the cue directs attention implicitly by being presented at the target location or explicitly by central directing cue such as an arrow; or the likelihood of cue-target location coherence is altered such that the cued location could become associated with the opposite target location.

The paradigms effects are typically assessed by reaction time or error rates but since multiple, possibly parallel, processes must take place between visual stimulus and motor response, we need more than behavioural data to disentangle the time course of the behavioural profile. EEG's high temporal definition, and increasing spatial capabilities, makes it a useful tool to delineate the processes involved in such a task. Several EEG studies have already been conducted using variations of the task (see (105) for a review of studies prior to 2006; (106–109)). The issue is that the EEG studies conducted have only explored one or two factor manipulations of the paradigm and subsequently used inaccurate statistics to test the results. Consider that the most recent ERP study of the IOR effect was one of the first to examine cue-locked ERPs and was considered novel in its approach by examining three different time windows of 10 author-selected components of interest from ERP (109)[*]. Statistically, the authors employed multiple one-way repeated measures ANOVAs (with no explicit control for multiple-testing), for each component and XYZ location of the maximal amplitude with the time window as the only factor tested. Furthermore, source reconstructed maps were created for

[*] Published in the journal Brain Topography; impact factor (2010) of 3.288; a relatively high value for this fairly specialised journal.

each time window with no statistical analysis of the results presented[*].

Here, we aimed to assess several factors simultaneously using reaction time and accuracy measures as well as high-density EEG recording. The paradigm was as chosen to be as close to the original one set out; but able to accommodate the desired factors. Therefore, we take a more open exploratory analysis technique to a well-established and researched paradigm with the idea of reanalysing and confirming previously made hypotheses yet at the same time be open to the discovery of novel findings in the data-set made possible by valid statistical methods which remain highly sensitive to different types of variability in results.

4.3 Method

4.3.1 Behavioural Task

Fourteen, healthy, right-handed, male participants (mean age = 25.4, SE = 0.98), completed a modified version of the classic Posner paradigm . Participants were paid for their time, and gave written consent. Participants were seated comfortably in front of a 24" LCD monitor (60Hz), at a distance of approximately 50cm. The task was fully programmed in the Psychtoolbox extension (version 3; freely available at psychtoolbox.org) to MATLAB (110). As illustrated in Figure 15 below, participants began each trial in the task by fixating on a cross in the middle of the screen. Two empty squares then appeared to the left and right of the fixation cross at a distance of 40cm (viewing angle on average of 84.7°). In 20% of the

[*] The article is recommended as an illuminating example of how complex some EEG analyses can become and still lack in validity.

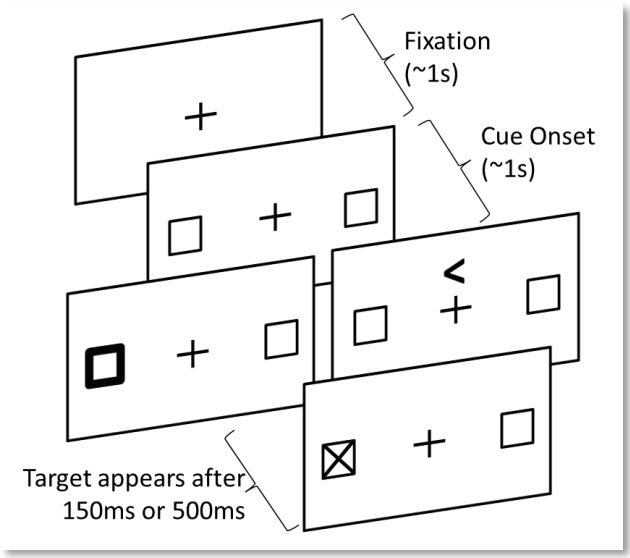
trials, deemed neutral trials, a target 'x' directly appeared in one of the two squares. However, in the majority of the trials, participants first saw a cue which could indicate where the target would subsequently appear. Valid cues, which accurately predicted the target location occurred on 75% of cued trials. The cue could take one of two forms: either the cue was presented externally, when the square's border would thicken; or internally with an arrow just above the fixation cross pointing to a square. For cued trials, the stimulus-onset-asynchrony (SOA), was either 150ms (short) or 500 ms (long). 400 total trials were presented to each participant, with a short break of 15-30 seconds after every 50 trials and a longer pause after 200 trials. All trials were pseudo-randomised in that the amount of each trial type, including left and right targets was fixed but the order of presentation was randomised for each participant.

4.3.2 EEG Recording and Analysis

EEG was measured from each participant using a 125-channel recording net in a geodesic arrangement. Signals were amplified and digitally sampled at 5000Hz using the Brainamp DC / MR amplifiers (BrainProducts). Using Analyzer2 (Brainproducts), the data was bandwidth filtered between 0.7Hz and 40Hz, downsampled to 250Hz, and re-referenced to the average activity over all channels. EOG artifacts from blinking were corrected using the Graham and Coles algorithm by constructing virtual VEOG and HEOG channels using a combination of the frontal electrodes. [Three] participants had severe EKG artifacts, primarily over posterior electrodes. An independent-component-analysis was conducted over the entire dataset was used to find, and subtract, components most loaded with the EKG artefact. Any further artifacts were semi-automatically marked under strict criteria (maximum voltage step 25µV/ms; maximum allowed total difference of 75µV in 200ms; maximum and minimum amplitude of

+/- 150µV; low activity of 0.5µV in 100ms). Those channels and time points were excluded from further analysis. Event-related potentials were calculated separately for both the cue-onset and the target-onset. Each ERP was baseline corrected using the mean activity of the period 200ms prior to the event.

Figure 15 – Basic overview of the Posner Task. Participants were required to keep their eyes on the fixation point at all times. External (thicker frame) or internal cues (an arrow), were presented to indicate the likely presence of the upcoming target (75% of cued trials). The target would appear either 150ms (short) or 500ms (long) after the cue. In 20% of all trials no cue was presented.



All channels and time-points were then statistically analysed using non-parametric methods following a threshold-free cluster-enhancement in order to enhance the statistical signal that were well supported by neighbouring channels and time-points. The default TFCE settings of E = 2/3, H = 2, were used as well as 2500 permutations to form the null distribution. The method has strict

control for multiple comparisons and therefore a threshold alpha value of 0.05 was used to determine significance.

4.4 Results

4.4.1 Behavioural: Reaction Times

Reaction times that were indicative of false starts or missed trails, under 100ms and above 1s, were marked as incorrect and subsequently left out of the reaction time analysis. Kolmogorov-Smirnov tests of normality were calculated for each dependent variable and all were found to be normally distributed (mean significance= 0.87, sd= 0.14). Reaction times for correct trials were submitted to a single repeated measures analysis of variance (ANOVA). The ANOVA was conducted with four within-subject factors: trial-validity (valid, invalid, and neutral); cue-location (external, internal); SOA (150,500); and target-side (left, right). Cue-dependent factors for neutral trials were included in the statistical analysis by design but ignored in further processing. Mauchly's Test of Sphericity found no significant deviations on any factor with the minimum Epsilon value found being 0.838. Results for the statistical analysis of each main effects are presented in Table 4.

Participants responded with a mean of 352ms (SE= 4ms). There was no interaction between all four, or any three of the factors. There were two separate interactions found between SOA and trial-validity ($F_{2,26}$= 4.756, p= 0.017) as well as SOA and cue location ($F_{1,13}$= 15.742, p= 0.002). For the SOA, trial-validity interaction, the main effects remained the case but it seemed as though that participants benefitted even further from longer SOAs in invalid trials (T_{13}= 2.341, p= 0.036; mean difference= 15ms, SE = 6ms). That is, the large difference between valid and invalidly cued trials for the short SOA was dramatically reduced. Although this is

not strictly the general finding of inhibition-of-return, it is consistent with the effect. For the SOA, cue-location interaction, main effects still hold true, but the benefits of a longer SOA seemed to be attenuated for externally presented cues (T_{13}= 4.022, p= 0.001; mean difference= 15ms, SE= 4ms).

Table 4 – Summary of each factor and level for reaction time (RT) and its standard error (SE); as well as behavioural accuracy (%). F-Values and there corresponding p-values are also given for each factor. Significant p-values are highlighted in bold.

Factor	Level	RT *(SE)*	F / *p*	%	F / *p*
Validity	Valid	318 *(4)*	33.280 ***>0.001***	98.33	13.55 ***0.003***
	Invalid	361 *(6)*		95.89	
	Neutral	376 *(7)*		99.82	
SOA	150	365 *(5)*	58.082 ***>0.001***	98.00	0.005 *0.942*
	500	339 *(4)*		98.04	
Cue Type	External	356 *(5)*	5.295 ***0.039***	98.08	0.06 *0.804*
	Internal	348 *(4)*		97.96	
Target Side	Left	362 *(4)*	14.478 ***0.002***	98.25	0.79 *0.389*
	Right	341 *(5)*		97.78	

4.4.2 Behavioural: Accuracy

Participants rarely committed errors (mean correct percentage= 98.0%, SE= 1.3%), which included false starts or late responses, not only incorrect target-side selection. Due to the nature of the variable, most factors showed significant deviations from sphericity, and although the Epsilon values remained fairly high, the lower-bound correction was taken as a conservative estimation.

For the significant main effect of trial validity for accuracy post-hoc Wilcoxon Signed Rank Tests indicated that all levels were significantly different from one another ($Cue_{Invalid}$ vs Cue_{Valid} : Z=

2.622, p= 0.009; $Cue_{Invalid}$ vs $Cue_{Neutral}$: Z= 2.938, p= 0.003; $Cue_{Neutral}$ vs Cue_{Valid} : Z= 2.271, p= 0.023).

Here, a three-way interaction was found between SOA, cue-location, and target-side ($F_{1,13}$= 6.195, p= 0.027). The largest difference we could find between the factors was, between left and right targets between externally (left side dominance of 2.3%, SE= 1.3%), and internally presented cues (right side dominance of 1.6%, SE= 0.7%), only at an SOA of 150ms. Another Wilcoxon Signed Ranks Test confirmed this difference (Z= 2.318, p= 0.02).

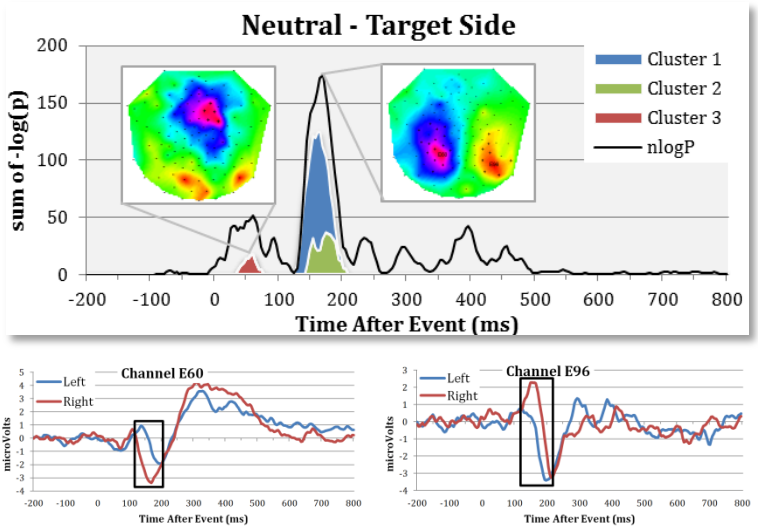
4.4.3 ERP: Target Side

In order to explore whether the highly significant behavioural differences for target-side were reflected in brain activity, this factor was examined on its own using the simple paired t-test version of the TFCE approach. The ERPs were taken only from neutral trials so that no cue events would contaminate the ERP of the target presentation. Figure 16 gives an overview of the results of this comparison.

Three separate clusters of significant results were found which highlight differences between the ERPs. The largest cluster spanned 190 channel-sample pairs which included 26 unique channels and ranged from 132ms to 200ms. The most significant point within the cluster occurred at 152ms over channel E66, located left-posterior (T_{13}= 11.69, p= 0.0004). Upon examination of the ERP the likely source of the difference was that targets presented to the right visual field had an earlier peak by 32 ms; moreover, the amplitude of the peak was also substantially larger by approximately 2 µV. The second largest cluster was essential the mirror-image of the first in terms of topography. It spanned 84 channel-sample pairs which included 13 unique channels and ranged from 144ms to 204ms. The peak in the cluster was found at

channel E95, right-posterior, at 176ms (T_{13}= -6.46, p= 0.0028). The source of this difference in the ERP was that targets presented to the right visual field showed a sharp positive deflection which was not apparent for left targets.

Figure 16 - Analysis results of differences between targets presented to the left versus to the right for neutral (no cue) trials. Clusters are calculated after the analysis and represent connected significant channel-sample pairs.



The third, smaller cluster of significant results spanned only 29 channel-sample pairs over just 6 channels from 36ms to 72ms after target presentation. The cluster was located in the central region, slightly right and anterior to the central electrode. This difference peaked at 56ms on channel E111 (T_{13}= 6.82, p= 0.0200). Examination of both ERPs indicated that these early differences were the result of differences in signal structure. Left targets showed a clear negative potential at 160ms whereas for targets presented on the right, this potential was much present earlier but much more varied in latency across subjects. Thus average ERPs

showed an early broad potential for right targets and a sharper potential at a later latency for left targets, resulting in the early significant differences.

The symmetric topographies would suggest equal but opposite differences for left and right targets. However, targets presented to the right visual field, left hemisphere, showed earlier latencies and stronger amplitude ERPs over left-sided electrodes; as well as an additional potential, not present for left-targets, over right-sided electrodes. These results are consistent with the right-target advantage for reaction times.

4.4.4 ERP: Cue Location vs. Target Side for SOA of 500ms

The previous analysis showed that presentation of the target to the left or right side showed significant differences in individual ERPs. To further examine whether a similar effect could be found for presentation of the cue, a 2x2 TFCE-ANOVA was used to assess the main effects of cue-location (external or internal) and target side (left or right), as well as any interaction between the factors.

For the main effect of cue location, analysis found three clusters of significant results. The largest significant cluster spanned 824 channel-sample pairs over 65 unique channels and ranged from 268ms to 468ms. Its peak significant point occurred at channel E91, a right posterior parietal channel at 340ms ($F_{1,13}$= 73.17, p= 0.0014). Examination of the corresponding ERP showed that internal cues resulted in an evoked-potential which was not present for external cues. The second largest cluster spanned 232 channel-sample pairs over 35 unique channels and ranged from 148ms to 212ms. The highest value in that cluster occurred at channel E13, an anterior-central channel at 168ms ($F_{1,13}$= 67.03, p= 0.0026). The difference seemed to be caused by a latency shift with internal cues processed

approximately 30 ms earlier than external cues. The third cluster of significant points spanned 83 channel-time pairs over just 13 unique channels and had a time range similar to the previous cluster from 152ms to 204ms. The peak significant point was found at channel E50, left posterior-occipital, 172ms after the cue presentation ($F_{1,13}$= 117.26, p= 0.0012). The values here reflected both differences in latency and amplitude of a negative deflecting ERP. Internally presented cues showed reduced latency and significantly higher amplitude (approximately 50ms earlier and 1.5 μV more negative).

Target side did not show any significant differences prior to the onset of actual target at 500ms. After the presentation of the target, whether the target was left or right showed a significant effect with almost identical characteristics as described for neutral trials (see 4.4.3 These two factors showed no points of interaction.

With no interaction effect, or specific target side effect directly after the presentation of the cue, we can conclude that there is no evidence for preferential processing for cues to either side. Thus the target side effects described for reaction time and neutral ERPs occur only at the time that the motor response is actually required. The three-way interaction of SOA, cue-location and target side for reaction time however remains unaccounted for by the ERP findings.

The reaction time differences for internal and external cues can be accounted for by these ERP findings; specifically, the significant interaction between SOA and cue-location for reaction times. Here we found earlier latencies with stronger amplitudes for internally-presented cues which fit well with overall better performance for those types of cues in participants' reaction times. Moreover, the significant differences between cue-types occurred after around or after 150ms; the time at which the target would

have appeared for the shorter SOAs. Therefore, the longer SOAs of 500ms allowed for the full processing of these two cue conditions which lead to increased behavioural differences specifically for that longer SOA.

4.4.5 ERP: SOA vs Trial Validity

Reaction time measures indicated a significant interaction between SOA and trial-validity, consistent with finding of inhibition-of-return even though no explicit reorientation of attention had taken place. In order to investigate the possible neurophysiological underpinnings of this effect, a 2x2 TFCE-ANOVA analysed both these factors and their interaction for target-locked responses since a trials validity is only determined at target presentation. The main effect of SOA is however essentially uninterpretable as the target-locked ERP would have its baseline at different times in respect to the cue event. As predicted then, the main effect of SOA began to show differences from 0ms to about 500ms post event when it is reasonable to assume the cue-locked activity specific to SOA had ceased.

A single significant cluster was found for differences in trial validity spanning 330 channel-sample pairs over 41 unique channels and ranging from 184ms to 300ms. The cluster was primarily located over right central channels and had its most significant peak at channel E103 232ms after the target appeared ($F_{1,13}$= 68.70, p= 0.0024). The difference reflected an earlier ERP latency for valid trials of approximately 40ms. A latency shift this late in the ERP would indicate some change of attention had already taken place for the invalid trials, but this was not reflected in any earlier significant differences for this comparison.

Both SOA and trial validity showed a significant interaction for target-locked ERPs. A single significant cluster of results was

found spanning 95 channel-sample pairs ranging from 308ms to 350ms. The clusters peak was located at channel E104, right-central, at 312ms ($F_{1,13}$= 60.03, p= 0.0056). The interaction probably stemmed from two distinct interactions in the data. The first interaction was likely that only the ERP response for invalid cues presented at an SOA of 150 had a pronounced delay of approximately 40 seconds, compared to the three other ERPs in the design. The second possible source for the interaction effect was that in the latter portion of the significant cluster, validly cued targets presented with an SOA of 500ms had reduced peak amplitude of approximately 1 µV in comparison to the three other ERPs.

4.5 Discussion

Although it is not the purpose here to discuss these results in the larger context of visual attention; the findings are indicative of real differences that have not yet been explored in other studies. The reason, at least in part, may be due to a lack of appropriate statistical methods. In particular, the significant results for interactions between factors show that various serial analyses of factors on their own will lead to misinterpretations of the data. Even though the current model only supports the analysis of two factors we were able to explore multiple factors (albeit with multiple tests), in order to provide new evidence. Importantly, despite the increased complexity, the results can still be explored in a fundamentally understandable way without having to resort to biased techniques.

Chapter 5 - Conclusion and Future Perspectives

5.1 Valid

A method's validity is its most crucial feature. Any major flaw in a method's theoretical underpinnings and one is able to completely discredit the results without ever even having to see them. With a solid approach that leaves little room for manipulation or bias, the output of the procedure can be trusted. Thus any critique of the research hypotheses will have to be related to the actual results themselves or their interpretation.

The TFCE method ensures its validity by analysing all data points available in the ERP and weighing their importance without bias. There is therefore no need to speculate about possible effects prior to recording the data, and limit the scope of the results. TFCE also avoids having to make several assumptions about the data's distribution and structure by using the permutation approach for inferential statistics; assumptions which are increasingly unlikely to be met in modern EEG datasets with a multitude of recorded channels and time points of potential differences. Although in the case of more complex designs, there is an increase in the flexibility of the approach, this is no more than any other approach when confronted with multiple factors. Moreover, TFCE avoids the use of arbitrary settings such as a cluster-forming-threshold, or definitions of channel neighbours which are possible sources of bias in the procedure. In order to run an analysis all that is required is the ERP data itself, and sufficient information about the location of the channels measured.

5.2 Sensitive

A valid method may guarantee that the results should at least be explored by others, but if the method is not sensitive enough to differences then there won't be any result to discuss. Therefore maximising a method's sensitivity, while maintaining a high standard of validity, is the logical next step in the process. The most used permutation method for EEG data has been the maximum-statistic approach; which has been shown to have the strongest control over the false positive rate, but also is the least sensitive. The TFCE approach also uses a maximum-statistic approach but enhances the data prior to inference testing by including information from neighbouring clusters to increase overall sensitivity.

It is nearly impossible to predict what kind of signals will be present in a dataset prior to collecting the data since small changes in source location will heavily influence the scalp topography and signal intensity. Furthermore, even in just a single ERP dataset, there are likely to be different types of signal present since activity closely synchronised with the triggering event will tend to have a more focal pattern in time; whereas secondary activity related to the event will be more broadly distributed in time. Both the cluster size and mass approaches have been shown to have better sensitivity than the maximum-statistic approach but do so at a cost to method validity. Moreover, by setting a single threshold, the methods are generally only sensitive to a specific range of ERP signals. TFCE on the other hand uses balanced weighing parameters for information about a data point's intensity and neighbourhood and is therefore sensitive to the many kinds of signal types present in EEG data. Information about the surrounding clusters of data is automatically taken over the entire signal, not just at a single arbitrary threshold. At the end of the process intense focal signals

are enhanced to levels which make the TFCE values directly comparable to more dispersed signals.

Sensitivity can also be increased by including further information into the analysis. As we have seen, TFCE can handle complex designs where an additional factor in the experimental design may account for some of the random variance in the primary factor of interest. Likewise, fixed covariates such as age or gender may also reduce variation in the ERP. Inclusion of non-fixed covariates like the results from the behavioural portion of the experiment (e.g. reaction time), may also predict variation of the ERP and would therefore act to increase sensitivity to the underlying signal. Since the initial-statistics are based on the general linear model, it is a straightforward process to include these additional explanatory factors into the analysis.

5.3 Interpretable

From a purely theoretical stand-point, a method's validity and sensitivity are the only two crucial aspects. Yet if this were also true from a practical stand-point, then it seems unlikely that the conventional modes of analysis would still be so common. After all, the method is a tool which should turn a large set of data, into a clear collection of information. The results should provide a clear picture of the overall outcomes of the experiment, and at the same time be able to answer specific questions.

A feature of the TFCE approach is that the output matrix of p-values is identical in shape to the input matrix. This matrix can then be viewed from different perspectives in order to visualise the data. For example, in the case of a specific hypothesis about a channel or time point, the corrected p-values can simply be taken from the corresponding points in the matric. Or, for a more

exploratory analysis, the summation of the negative log of the p-values gives a timeline of significance over the entire ERP. Specific time samples can then be selected and scalp topographies displayed showing the localisation of the effect. By combining the timeline with topographies from peak significance points one could give a general overview of the entirety of the results in a single figure. Thus, even if the research question was regarding a specific feature of the ERP, it would still be worthwhile to conduct a full-analysis since this would: put the specific feature into perspective (e.g. onset, range, and peak of the feature); provide far more confidence in the effect when its corrected over all channels and time points; and allow the research community to explore other hypotheses using your results rather than having to collect entirely new data.

A further attribute of TFCE is that the resulting p-values actually represent the amount of trust we should place in that difference and need no extra qualification to be understood in context. Here, two p-values, even when they are representative of very different kinds of signals are directly comparable. Often with other methods, a p-value may be given with additional information about its actual intensity or the significance of surrounding structures. In SPM for example, it is commonplace to provide the significance value of a specific feature of the ERP in terms of its uncorrected, family-wise error corrected, false-discovery rate corrected, and/or cluster-level significance level. This can lead to data which is non-significant with some corrections being reported as significant because of its significance in some other measure (this most commonly happens with the uncorrected p-value being reported with some weak justification). With the TFCE approach, the inference statistics are calculated on data which already incorporates all this information into a single value which can be then directly reported.

5.4 Methodological Limitations

What follows is a description of three theoretical drawbacks to the TFCE method which are described below. These limitations are considered inherent to the procedure and hence their effects should be understood and accepted. There are further limitations of the procedure which are not inherent, and can be improved upon which are discussed in the later section 5.5

5.4.1 Bipolar deflections of a single source

Apparent in almost every EEG dataset is that for every positive deflection found, there is a corresponding negative deflection. This is especially apparent for signals created from a single dipole (see section 2.3 for simulated dipoles and their scalp topographies). Therefore, the topography of a single source signal will usually be defined by two clusters of data; one positive and another negative. The TFCE approach does not take this information into account and treats each cluster of results as independent evidence for a significant signal source. It may therefore be argued that any method which involves improving sensitivity through data clustering is ultimately flawed and should not be used.

Clearly, we feel that this strong opinion is unwarranted; especially given the results presented for both simulated and real datasets of varying designs and underlying physiology. Furthermore, this argument could be considered a fundamental issue in any EEG analysis method; especially to those methods that do not examine the entire dataset for significance. What is required is caution when interpreting significant clusters as representative of significant activity in the cortical area underlying the significant effect on that channel-sample pair. A tentative solution to this issue may be to first estimate the underlying source activity and then perform a

TFCE analysis on this larger 4-dimensional dataset (for a discussion on benefits and pitfalls of analysis on source reconstructed data see section 6.4).

5.4.2 Reference dependent

A crucial step to any EEG recording and subsequent analysis is the definition of the reference to which all activity is compared to. All potentials found after pre-processing are by definition potentials in relation to some reference. Historically, one of the most common references has been to use the activity over the mastoid bone just behind the ear. More recently, with the increase in the number of electrodes, each electrode is referenced to the average activity over all recorded electrodes, which should summate to approximately zero. Although arguments can, and have been made for and against the use of certain electrode montages, the shape of ERP-waves is highly reference dependent. That is, depending on the reference electrode(s) chosen, the shape and intensity of individual channel waveforms can change completely. For example, alpha activity (rhythmic 8-12Hz), can be seen over all electrodes using mastoid references but only over occipital sites for average reference montages; this is because the alpha activity is actually being measured at the mastoid reference as well and then projected by differentiation to all sites which are then referenced to it. Importantly the overall shape of the topography does not change and selecting another reference is akin to changing the sea level, while the underlying landscape remains constant (this is the key feature and argument for proponents of microstate analysis described in section 1.2

Imagine for instance a simple topography consisting of a single peak (or hill), and a completely flat baseline topography. Shifting the distance between the two topographies could result in

a difference of just the peak when the flat parts are aligned, or a difference of almost the entire field except for the peak if the baseline is shifted upwards. For most waveform analysis that either look at just the peak differences between the waveforms (maximum-statistic), or how large the distribution of differences is (cluster-size), a different reference can completely 'make or break' the result structure.

For TFCE, this is far less dramatic since both differences in waveform intensities and cluster sizes are looked at. Such that, if the values of E and H are appropriately configured to take the best of both topographical features, any loss of intensity differences, by varying the reference, is compensated for to some extent by some proportional increase in cluster sizes. In other words, references that eliminate differences in one channel will inevitably create proportional differences in other channels. Moreover, the use of any other reference montage other than the 'average-reference' has become increasingly uncommon; even the proponents of the topographical method, argue for the use of this reference procedure for their actual statistical waveform analyses (31, 32, 111). With a sufficient number of electrodes the average reference has been shown to be a good approximation to the zero voltage line, and hence, under those conditions, a relatively good approximation to reference-free measurements whilst still retaining information on relative strength of the signal (112). Thus, although the TFCE method is affected by its reference dependence, the effect has far less of an impact on the statistical analysis and more on the inferences that can be made about specific differences at certain channels.

5.4.3 Recording parameters influence cluster sizes

For fMRI data, the structure of the input is relatively fixed in terms of image dimensions. Data from ERPs can vary widely in terms of the number of channels recorded and analysed; the time range included in the analysis, both baseline and after the event; and the sampling rate of the ERP. Cluster extent is calculated as an absolute value from the input. Thus, if our data was down-sampled to 500Hz rather than 250Hz, or if we used a 256 channel camp rather than 128 channels, we would ultimately obtain cluster-sizes twice as large.

This may, at first glance, seem like a major flaw in the analysis process; however there are two reasons why it has little effect on the results, not only in terms of validity but also the actual obtained significance values. Although the actual TFCE values may be altered, the shape of the data remains unchanged. An increase or decrease in the sampling rate scales all cluster sizes equally. Thus, we essentially just multiply the TFCE values by a scalar. Secondly, because we base our inferences on permutations of the same scaled dataset, the cluster sizes of the permutated datasets will also be multiplied by the same scalar. Thus, our original TFCE values and permuted TFCE values still maintain the same relationship to each other and the p-values would be kept constant[*]. Clearly, sampling rate will have a larger influence on the analysis results once it changes the shape of the data at quite low rates. Yet this is the case with every analysis method, and is generally avoided in the pre-processing stage. In any case, we recommend a sampling rate that is several multiples of the highest

[*] This was empirically tested on simulated source 6 by up-sampling the original dataset by double its original sampling rate. Calculated p-values showed only minor variation due to the randomisation of permutation process.

frequency inspected in the data. For most standard EEG datasets this minimum sampling rate is likely to be around 200Hz.

5.5 Future work

We believe the TFCE approach is the most optimal method currently available for analysis of EEG datasets; in particular for open discovery of significant differences in large datasets when a-priori hypothesis are untenable. Moreover, the TFCE approach is designed with large datasets in mind; common to modern research in the field. However, there are still several aspects of the process which have room for improvement or potential ideas that require further investigation.

5.5.1 Initial-statistics

The t-statistic is automatically taken as the default initial statistic and as a result the open choice of initial-statistic is often overlooked in analyses. Although the t-test may be the most common and optimal measure of differences between two datasets, there may nonetheless exist further measures which may be more of interest to the EEG researcher. Since the permutation approach empirically calculates the null distribution from the data, it is open to whichever measure of differences the experimenter chooses; without any necessary alterations to other aspects of the process.

Currently implemented in the algorithm is the possibility to directly calculate the differences in either means or variance between the two datasets in question. This was implemented because the t-test is sensitive to both types of differences and a significant result without further exploration may be attributed purely to differences in variability when the means could be

identical. Thus, using only the direct measure, one would be able to unequivocally determine whether significant comparisons were due to either differences in means or variance, or both.

Other measures have also been used in previous research (such as Hotellings T2 (75)). These may well be worth investigating as to whether they provide more accurate results for EEG datasets. Caution must be taken when examining several options for the initial-statistic in that it may provide too much flexibility for the researcher to experiment with, leading to an overall increase in false positives. In the end clear guidelines must be determined before allowing a truly open choice of initial-statistics. Ideally the choice will be determined by the data itself, which would allow for an automatic selection of optimal statistic by some algorithm.

A further expansion to the initial-statistics concerns the way the statistics are calculated in the first place. The first improvement that can be made is, rather than only use the variance from each channel-sample pair to calculate statistics, the variance can be pooled to give a more accurate estimate of the parameter. That is, each channel-sample pair's variance is taken partly from its own variance and partly from its neighbouring channels and time points. This will have the effect of reducing the influence of special artefacts in one of the contributing datasets (e.g. a single, quite variable participant) for a specific channel-sample. Furthermore, sphericity (the equality of the differences in variance), will hold true for the dataset as a whole but not necessarily for individual data-points. Thus, pooling variance will improve sphericity in the dataset and create more accurate statistics that are more representative of the real differences in the data.

The second overall improvement would be to automatically calculate group average statistics. That is, even when looking for the

differences between conditions or groups, it would be good to already know which channel-sample pairs are significantly different from the baseline zero. In other words, which parts of the ERP are actual potentials and which peaks and valleys are just random noise left from an imperfect averaging process. This could then be used to further interpretation of the subsequent results for conditions or groups.

5.5.2 Expansion of designs

Currently, the TFCE algorithm is able to efficiently handle any single factor design using the appropriate t-tests for two groups or conditions; in the case of multiple groups, a one-way ANOVA; or a repeated measures design. The previous chapter demonstrated the method is also capable of handling two factors simultaneously, both repeated measures. Furthermore, scripts have been written and informally tested which can analyse any type of two factor design; whether both factors are independent observations, or a mixed design. However, those are currently limited to designs with the same number of participants for all groups[*].

Work is already in progress which extends the analysis principles to include designs with an unbalanced number of subjects per level by using weighted means in the ANOVA. In addition, n-factor designs may be run using a beta-version but the generalisation process to more than two factors currently comes at a considerable loss of processing speed. Concurrently, only permutation of the raw data is possible; however as argued in section Chapter 4 permutation of the residuals in the linear model

[*] As weighted averages take a considerable amount more time to calculate, thus dramatically reducing the efficiency of the entire permutation process. It should be noted though that repeated measures analysis will always have an equal number of observations by definition.

may be a more valid and sensitive approach. Therefore, the option to select permutation method for more complex designs should be explored.

Future implementations may also include non-parametric versions of the initial-statistics tests such as the Mann-Whitney-U test for two sample designs; the Kruskal-Wallis test for an equivalent to the one-way ANOVA design; or the Friedman test for more complex designs without the need for approximation of the interaction term. Since inferences are made by the permutation approach, the non-parametric test versions do not imply that the t-tests and ANOVA versions are invalid. Rather, they should be used to produce further alternatives to the initial-statistic should the researcher feel that those tests do not accurately represent the data. As with the warning for initial-statistics, alternative approaches should be a free choice to the user, but rather automatically taken by the algorithm once certain aspects of the data are calculated. The exact decision tree necessary should be thoroughly justified both theoretically and empirically using simulated data.

5.5.3 Data smoothing

Smoothing raw data prior to analysis is fairly common practice in neuroimaging, especially in fMRI analysis. A 3D guassian kernel of a particular size is usually used to smooth the fMRI image. This will have the effect of improving SNR by reducing the impact of randomly distributed activations due to noise. In the introductory paper on TFCE (81), it was shown that although the implementation of TFCE on the raw image performed fairly well, when combined with a smoothing technique the results were dramatically improved.

Smoothing an image can be seen to be an equivalent to using a low-pass filter on the temporal domain of EEG. In section 2.6 we demonstrated improvements in signal detection after using temporal filters on simulated EEG data; but currently there is no easy method to also smooth data in the spatial domain. SPM is capable of spatial smoothing because they use interpolate the EEG data to form continuous 3D images. The drawbacks of this interpolation are discussed later in 6.3 but suffice to say that the costs of the procedure are too high to be a viable option. However, given the way TFCE calculates neighbouring channels (described in 6.3.1 the information could also be used to smooth data in the spatial domain as well as in the temporal domain. Given the improved sensitivity of smoothing reported in the original article, this option deserves exploration in future work. The benefits might be especially visible in multi-subject studies where spatial smoothing may adjust for variation in EEG topography caused by different head sizes, head conductivity variability, and inaccuracies in electrode positioning.

5.5.4 Nonstationarity

Processes whose statistical properties (e.g. mean, variance, correlation), are subject to change are referred to as nonstationary. For time series nonstationarity is seen as trends, cycles or random walks of the data. Spatial nonstationarity can be seen as a non-uniform smoothness in the data. Nonstationarity is a problem in data because many statistical algorithms have an implicit assumption of stationarity, and deviations, like deviations from any assumption, will lead to biases. The effects have been largely examined for fMRI analysis and have been corrections proposed (113, 80), including the TFCE for MRI approach (82), but to the best of our knowledge, no such literature exists for EEG datasets. For EEG, if the smoothness of the spatial topography differs over the

scalp, then larger cluster sizes would be expected for those areas by chance alone. Thus, future work should attempt to find estimates for data stationarity, especially in the spatial domain, in order to reduce any of these potential biases.

5.5.5 Optimal signal detection assessment

Given that one of the goals of the TFCE method is to eliminate post-hoc user qualification of the results of an analysis, it is somewhat hypocritical that such qualifications were necessary in order to further explain the results of the assessment measures for methods for simulated signals. For example, it was argued that the default parameters settings for TFCE should despite other parameters showing superior performance because of the limitations of the comparisons measures. Section 0already argued why the chosen measures of precision, recall, and MCC were nonetheless the optimal choices for comparison, given the major limitations of other measures. However, this is not to say that these measures are the best possible measures conceivable. Just as the proper analysis of real EEG datasets is crucial to the scientific community, the ability to accurately compare the different methods of analysis is essential to the achievement of that goal.

Thus future work should look into providing a signal detection method that is able to handle 'fuzzy' truth (i.e. degrees of signal intensity as opposed to binary ground truth of signal/no signal); is not biased by imbalanced datasets that contain far more true negatives than true positives (as would be the case with most simulated datasets for EEG); one that gives a fair assessment for different signal types (as opposed to being biased to large clusters which contain more true positive signals by definition); and finally, one that would have the ability to assess the result structures of a range of significance levels from zero to a given threshold.

5.5.6 Software development

Currently, the TFCE algorithm is implemented in two user-friendly programs. The first is a simple menu asking the user to locate the ERP datasets files on the computer, as well as the electrode coordinates which then continues to run the analysis automatically. The second is the result viewer which was examined in some detail in section 2.2.2 There are already several software solutions for basic pre-processing of EEG data, some even freely available such as EEGLAB or SPM (free if one already has access to MATLAB). Integration of the current TFCE tools into those programs would allow for a complete analysis of EEG datasets from raw recorded data to interpretable results and would thus increase the likelihood that researchers use the method. Work has already begun to implement current algorithms into EEGLAB (because of its pre-processing capabilities, and ability to read and write to all major EEG data types), as well as Brainstorm (because of its intuitive data structure and visualisation capabilities; as well as its possibilities of integration of data from other modalities such as MRI).

However, given that no single program has optimal features for pre-processing and visualisation, it may well be worth exploring the possibility of a full-standalone program for analysis. This option would provide two major advantages. Firstly, the program would be entirely run without the necessity for MATLAB, which would open its use for researchers without an available license (a possible issue for clinicians in smaller hospitals), and could substantially improve the runtime of the algorithm. Secondly, the program would be developed specifically with the features of TFCE in mind, and design of the interface could maximise the available features, and eliminate the use of redundant analysis options.

Chapter 6 - Appendix

6.1 How many permutations are sufficient?

Even for just 10 participants per group there would be 184'756 ways in which we could permute the labels and create new datasets. This value increases exponentially as we increase the amount of labels (50 trials per condition in a single subject study could be permuted in the order of 10^{29}). This mathematical fact makes it necessary to reduce the actual amount of permutations calculated in practice. This reduction should still allow a final p-value within an acceptable range of the exact value but should be as minimal as possible to keep processing times to a reasonable range. For this reason, in practice most permutation tests are actually approximations to the exact test. This should not be confused however with approximate tests, such as all parametric variations, which rely on several assumptions to approximate to the exact result.

In order to empirically assess how many permutations would be necessary and sufficient we generated random data (using MATLAB), and created 7 different datasets corresponding to 8, 10, 12, 15, 20, 30, and 50 labels in 2 groups/conditions. The p-value was calculated for these datasets 50 times in order to obtain a mean p-value, and more crucially, a standard deviation and range for the p-values in order to understand to what degree of accuracy the p-value could be taken from any of the 50 tests conducted. This test was performed using 9 values for the number of permutations, 100, 250, 500, 1000, 2500, 5000, 10'000, 25'000, and 50'000.

All datasets started with a relatively high mean p-value (in comparison to the calculated exact values for 8, 10, and 12

datasets, and the average mean for the other datasets where exact values could not be calculated). All p-values then steadily decreased and by 1000 permutations began to converge around the exact value. This has an important consequence, with few permutations, p-values tend to be higher than their real value, and as such using too few will tend to produce results which are too conservative.

As expected, the standard deviation of the p-values steadily decreases with an increase in the number of permutations. Remarkably the absolute standard deviation seems to be independent of the total number of possible permutations. Thus, an ideal number of permutations, from the perspective of the standard deviation, depends on what degree of accuracy one is willing to accept. However, after about 2000 permutations the gain in accuracy steeply declines. Moreover, the increase in the number of permutations is directly proportional to the computational time and thus, after about 2000 permutations, for a gain in accuracy of less than 0.005 we would see a 10-fold increase in the time needed to analyse the data. In conclusion, any number of permutations above 2000 is likely to give a fairly accurate result, although using more is always recommended if the computational resources allow it since the variance between analyses continues to systematically decrease above this value. Moreover, since too few permutations results in conservativeness of the test, if borderline significance values are found, it may well be worth using an increased number of permutations in order to determine whether that calculated value is actually significant when the p-value nears its exact rate.

6.2 Analysis method pseudo-code

The actual programming code which runs the TFCE analysis, which have both m-file scripts from MATLAB and c-file scripts which

require mex-compilation, is open source and will soon be made publicly available at no cost. What follows below is referred to as pseudo-code. This is in attempt to show the reader the inner workings of the program without the need to understand formal programming language. Thus, the linearity of the program is displayed but with a balance of normal, conversational language in the hopes the algorithm becomes understandable. Comments are presented in brackets and italicised).

>> Load data into the Matlab workspace and check for consistencies
>> Calculate channel neighbours *(see section 6.3*
>> Calculate observed t-values *(paired or unpaired)*
>> Run TFCE calculation on positive and negative t-values separately[*] *(see 6.2.1 for pseudo-code of TFCE values),* then recombine values
>> FOR i = 1 : number of permutations
>> Create randomised dataset *(for independent t-test this is done by shuffling the participants; for paired t-test data is multiplied randomly by 1 or negative 1)*
>> Calculate t-values for randomised dataset
>> Run TFCE calculation on positive and negative values separately
>> Find the maximum absolute TFCE value and store
>> END
>> Find each observed TFCE value in histogram of maximum values and calculate proportion of more extreme values to obtain p-values

[*] For EEG positive differences are just as likely to occur as negative values and so TFCE is performed for positive t-values and negative values t-values separately (by setting the other to zero), and later re-combined into a single dataset for permutation thresholding. This is preferable to using absolute values since it avoids the possibility that positive and negative differences which are spatial neighbours are seen as part of the same cluster. For EEG setups using relatively few channels, even large differences may be seen as spatial neighbours and using absolute differences could substantially bias the results.

6.2.1 TFCE calculation pseudo-code

```
>> Find maximum t-value in input data
>> Calculate thresholding steps from 0 to maximum t-value
>> FOR i = 1 : number of thresholding steps (50 is default)
>>    FOR j = 1 : number of data-points (channel x sample)
>>       IF data-point is over threshold
>>          Look at channel neighbours and time points for other channel-
            sample pairs that are also over that threshold
>>          Multiply all found data-points in cluster by TFCE equation
>>       END
>>    END
>> END
```

6.3 Calculating neighbours

In order to calculate any sort of cluster, the idea of adjacent points in data must be well defined. In the case of fMRI data, a certain voxels' neighbours can be easily defined as the 6, 18, or 26 voxels surrounding it depending on whether one considers faces, edges or corners as neighbours. In EEG data, time and frequency samples also have clear neighbouring points. Channels on the other hand are sparse samplings of a 2D surface, the scalp, in 3D space. Therefore, channels are not consistently organised into a neat grid and determining a channels neighbours is a non-trivial problem.

As described in section 1.5 SPM, although well known for its analyses of MRI data, can also be used to analyse EEG data. The statistical process it takes is essentially the same as for fMRI data as the EEG time-space data is converted into activation maps using the 2D topography maps generated after interpolation of the channels collected. A 3D image is the generated by stacking the 2D maps over

consecutive time points. Projection of the 3D locations onto a 2D grid removes a dimension of data, and interpolation of channels over a uniform grid makes calculating neighbouring data points a matter of searching adjacent pixels. Although a seemingly elegant solution, there are several issues with this approach. Immediately apparent is that the process of interpolation will actually result in more data points than we originally recorded, and in doing so only adds to the multiple comparisons problem, as well increasing computational time. Since the data from any number of channels is made into the same image size, the benefit of recording a large number of channels is dramatically reduced (although still maintained in the analyses through the degrees of freedom). Data from 10 channels or 256 channels would still interpolate to build a 32x32 or 64x64 pixel image for each sample.

Secondly, we have lost specificity in our electrode location by projecting them onto a 2D surface, usually by assuming a constant head radius. Although this may only seem like a minor precision issue, 2D projections, like when creating large maps of the earth's surface, cannot maintain distance or area relationships. Thus, when interpolating data between channels we will obtain larger or smaller clusters solely depending on the distance biases in the 2D projection. Furthermore, for non-uniform electrode coordinates where a single channel is relatively alone, the interpolation would result in a disproportionately large are representing a single channel. Conversely, a dense area of electrodes, all showing significant results would still result in a fairly small cluster despite having a lot of supporting information.

Maris used a different approach implemented in the software Fieldtrip (114), and classified channels as neighbours if they were within a 4 cm radius (75). Despite being arbitrary, such an absolute value poses problems for the varying amounts of electrodes that can be measured. Not only will there be no clusters

if the electrodes are more than 4 cm apart, but a further problem arises if there are more than 2 channels in a space of 4 cm. For example, if two electrodes, 4 cm apart, showed significant differences but between them there was a non-significant electrode, then it stands to reason that the two significant channels should not form a cluster. Cleary, this value of 4 cm can be altered for different electrode configurations, but it is both tedious and still arbitrary to define a new optimal neighbouring distance for subject and analysis.

6.3.1 Triangulation of Electrode Coordinates

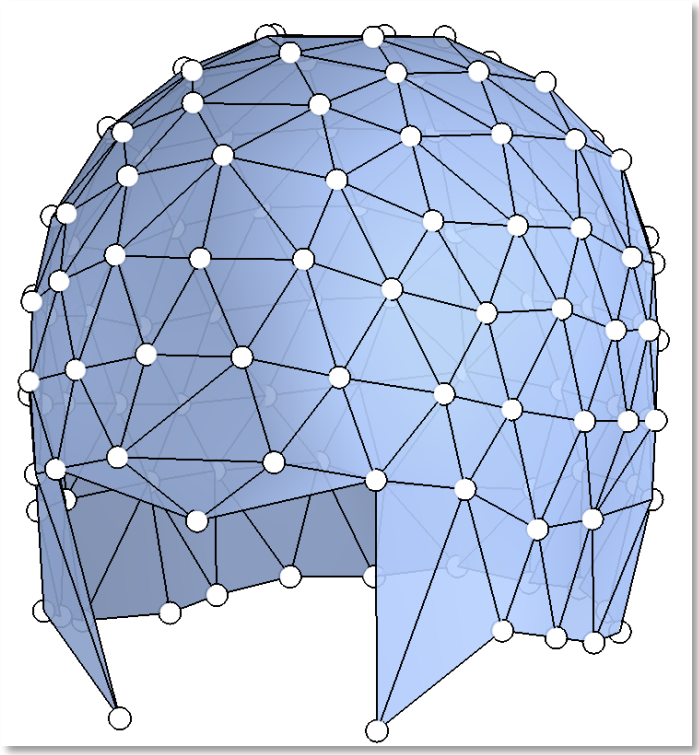
Here we introduce what we believe is a novel method of calculating channel neighbours using only the 3D coordinates of the sensors as input. Ideally, just as with the choice of statistic, there should be no arbitrary user parameter newly selected for each analysis, and the definition of neighbouring channels should be stable. However, the calculation should be flexible across a wide range of channel positions and total number of channels. Furthermore, we should be able to use the precise electrode coordinates in 3D space when they are recorded to increase sensitivity for datasets from a single participant, or averaged electrode coordinates for group studies.

The first step is to calculate the convex hull of the coordinates in 3D space. This is essentially calculating the triangles which would join all the channels to form a solid, empty polyhedron. The objective then is to automatically remove unwanted triangles such as the ones which would connect the most anterior channels to the most posterior ones (through the head), by searching the triangles for precisely the ones connecting three channels on the outer boundary of the electrode array. Once this has been done the mean triangle perimeters are calculated, and triangles are removed that have perimeters longer than 3 times the

standard deviation of that mean as this will remove perimeters that lie outside 99.73% of the normal, a highly conservative figure. This has the effect of removing triangles which are far larger than the mean, hence removing triangles between channels with large separations, such as the ones above and below the ear, or channels across removed channels.

The algorithm which accomplishes these previous steps has been adapted from the Brainstorm program (115), which uses the calculated triangles solely for displaying electrode arrays. Finally, for each channel a list of neighbouring channels is created by finding the channels which share edges with the channel in question. The channel neighbours found can be inspected visually or as a table and any extreme irregularities could be changed manually before further calculation. This method is subsequently used for TFCE calculations as well as cluster size and mass. Thus making all cluster methods used here already superior in this respect to those used elsewhere. Figure 17 shows is an example of the results.

Figure 17 - Triangulation of neighbouring channels for a 129 channel cap in a geodesic array



6.4 Analysis of source reconstructed data

It has become common practice to first use source analysis on the EEG data and then compute statistics based on the computed sources (116–118). Although it would be technically possible to run a TFCE enhancement on the 3D source data over time, we do not recommend this approach for three reasons. Firstly, we are currently far from a standard method of source reconstruction and we would therefore open the data to a new set

of possible biases. Currently, source reconstructions with minimal assumptions tend to either produce large unspecific sources, but an increase in specificity requires an increase in, often not empirically justified, assumptions about its location. Secondly, even if these assumptions were perfect, small errors or artefacts in the scalp data will lead to much larger variation in the source localisation (119). Therefore, non-significant differences from random variation in the data may seem significant once source analysis has been carried out. Thirdly, the addition of a third spatial dimension to our data will dramatically increase the number of points we need to analyse and so increase the time necessary for analysis and increase our chance for false positives.

Thus with source analysis we may drastically reduce the integrity of our data with little or no benefit to our analysis since real differences in the source of the activity should also be evident in our actual recording on the scalp. This is not to say that source analysis is not a useful tool but rather that we see no real advantage to performing the statistical analysis after localisation. Rather, the results of the statistical analysis on the EEG data should act as justification for the time points chosen to perform various source localisations. This would then be an additional aid to visualise and interpret the findings.

Acknowledgments

This thesis and the other research projects I have been privileged to be a part of would not have been possible without the help and support from many people.

Primarily I would like to thank my supervising committee: Lutz Jäncke from the department of neuropsychology at the University of Zurich; Ramin Khatami from sleep department at the clinic Barmelweid; and Peter Brugger from the department of neurology at the University Hospital Zurich; for agreeing to look over my written thesis and giving me the opportunity to defend its contents. In my case it took more coordination than usual, and I am extremely thankful that through the dedication of Professor Jäncke and Dr. Khatami it all worked out.

I am especially grateful to Ramin Khatami for his almost daily willingness to discuss all aspects of whatever research project I was busy with that day and taking an extraordinary amount of time to listen and provide his advice. Any other supervisor would have told me to drop many of the ideas presented in this thesis, but his patience and trust in my ideas allowed me to pursue my interests and I hope that he feels that trust has paid off. Another special thanks goes to Peter Brugger who was willing to give me my first chance at doing an independent research project at the University Hospital in Zürich.

For both their academic assistance and their friendship I thank: Corinne Tamagni, Leonie Hilti, from the University Hospital in Zurich; Sonja Tartarotti at the clinic Barmelweid; and Marcel Nicklaus at the University of Bern.

I thank both the Swiss National Science Foundation (SNF) for my PhD funding and the clinic Barmelweid for funding the experimental research of many of the projects not mentioned in this thesis. Moreover, I thank the entire staff, especially of the sleep department, of clinic Barmelweid for creating such a warm and welcoming place to come to work every day.

For the work on TFCE I thank Tom Nichols for his in depth discussions on permutation statistics and TFCE; Christian Gaser for providing me with the programming scripts for TFCE for fMRI which formed the basis of my own scripts; and to Pau Coma for taking a lot of time to discuss the entire project and working through all the different ways which we could optimise the scripts.

On a more personal level I thank Eva Kuske and the entire Kuske family for becoming my second family over the past year and providing their support and seemingly unending kindness.

Lastly I thank my family; Henk, Lineke, and Eline Mensen. Everything I do is to make them proud of me, and they are a constant motivation to keep going especially if things aren't always easy. Also thanks to my extended family; Bosco, Charou, Flor, Socks, Samba and Groovy. And finally a very special thanks to the late Indy, Donya, and Mischa.

References

1. Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2:e124.

2. García-Berthou E, Alcaraz C (2004) Incongruence between test statistics and P values in medical papers. *BMC Med Res Methodol* 4:13.

3. Jeng M (2006) Error in statistical tests of error in statistical tests. *BMC Med Res Methodol* 6:45.

4. Bakker M, Wicherts JM (2011) The (mis)reporting of statistical results in psychology journals. *Behav Res Methods* 43:666–678.

5. Berle D, Starcevic V (2007) Inconsistencies between reported test statistics and p-values in two psychiatry journals. *Int J Methods Psychiatr Res* 16:202–207.

6. Wicherts JM, Bakker M, Molenaar D (2011) Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE* 6:e26828.

7. Sterne JA, Davey Smith G (2001) Sifting the evidence-what's wrong with significance tests? *BMJ* 322:226–231.

8. Fisher R (1925) *Statistical methods for research workers* (Oliver and Boyd, Edinburgh). 1st Ed.

9. Hayat MJ (2010) Understanding Statistical Significance. *Nursing Research* 59:219–223.

10. Bergmann TO et al. (2012) EEG-Guided Transcranial Magnetic Stimulation Reveals Rapid Shifts in Motor Cortical Excitability during the Human Sleep Slow Oscillation. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 32:243–253.

11. Silber B et al. (2011) The acute effects of d-amphetamine and d-methamphetamine on ERP components in humans. *European*

Neuropsychopharmacology: The Journal of the European College of Neuropsychopharmacology.

12. Moont R, Crispel Y, Lev R, Pud D, Yarnitsky D (2011) Temporal changes in cortical activation during distraction from pain: A comparative LORETA study with conditioned pain modulation. *Brain Research*.

13. Reed P, Savile A, Truzoli R (2012) Event related potential analysis of stimulus over-selectivity. *Res Dev Disabil* 33:655–662.

14. Conrad M, Recio G, Jacobs AM (2011) The Time Course of Emotion Effects in First and Second Language Processing: A Cross Cultural ERP Study with German–Spanish Bilinguals. *Front Psychol* 2.

15. Decicco JM, Solomon B, Dennis TA (2012) Neural Correlates of Cognitive Reappraisal in Children: An ERP Study. *Developmental Cognitive Neuroscience: A Journal for Cognitive, Affective and Social Developmental Neuroscience* 2:79–80.

16. Ortiz-Mantilla S, Hämäläinen JA, Benasich AA (2011) Time course of ERP generators to syllables in infants: A source localization study using age-appropriate brain templates. *NeuroImage*.

17. Nozaradan S, Peretz I, Mouraux A (2011) Steady-state evoked potentials as an index of multisensory temporal binding. *NeuroImage*.

18. Houdayer E et al. (2011) Movement preparation and cortical processing of afferent inputs in cortical tremor: An event-related (de)synchronization (ERD/ERS) study. *Clin Neurophysiol*.

19. Lee T-W, Yu YW-Y, Wu H-C, Chen T-J (2011) Do resting brain dynamics predict oddball evoked-potential? *BMC Neurosci* 12:121.

20. Bernat EM, Williams WJ, Gehring WJ (2005) Decomposing ERP time-frequency energy using PCA. *Clin Neurophysiol* 116:1314–1334.

21. Degabriele R, Lagopoulos J, Malhi G (2011) Neural correlates of emotional face processing in bipolar disorder: An event-related potential study. *J Affect Disord*.

22. Leleu A et al. (2010) Perceptual interactions between visual processing of facial familiarity and emotional expression: an event-related potentials study during task-switching. *Neurosci. Lett* 482:106–111.

23. Rossion B, Caharel S (2011) ERP evidence for the speed of face categorization in the human brain: Disentangling the contribution of low-level visual cues from face perception. *Vision Res*.

24. Quandt LC, Marshall PJ, Bouquet CA, Young T, Shipley TF (2011) Experience with novel actions modulates frontal alpha EEG desynchronization. *Neurosci Lett*.

25. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci* 12:535–540.

26. Picton TW et al. (2000) Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. *Psychophysiology* 37:127–152. Available at:

27. Johnson JS, Olshausen BA (2003) Timecourse of neural signatures of object recognition. *J Vis* 3:499–512.

28. Spencer KM, Dien J, Donchin E (1999) A componential analysis of the ERP elicited by novel events using a dense electrode array. *Psychophysiology* 36:409–414.

29. Spencer KM, Dien J, Donchin E (2001) Spatiotemporal analysis of the late ERP responses to deviant stimuli. *Psychophysiology* 38:343–358.

30. Ferree TC, Brier MR, Hart J Jr, Kraut MA (2009) Space-time-frequency analysis of EEG data using within-subject statistical tests followed by sequential PCA. *Neuroimage* 45:109–121.

31. Murray MM, Brunet D, Michel CM (2008) Topographic ERP analyses: a step-by-step tutorial review. *Brain Topogr* 20:249–264.

32. Brunet D, Murray MM, Michel CM (2011) Spatiotemporal analysis of multichannel EEG: CARTOOL. *Comput Intell Neurosci* 2011:813870.

33. Guthrie D, Buchwald JS (1991) Significance testing of difference potentials. *Psychophysiology* 28:240–244.

34. Murray MM et al. (2004) Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging. *Neuroimage* 21:125–135.

35. De Santis L, Spierer L, Clarke S, Murray MM (2007) Getting in touch: segregated somatosensory what and where pathways in humans revealed by electrical neuroimaging. *Neuroimage* 37:890–903.

36. Goodman SN (1998) Multiple comparisons, explained. *Am. J. Epidemiol.* 147:807–812; discussion 815.

37. Kiebel S, Friston F (2004) Statistical parametric mapping for event-related potentials (II): a hierarchical temporal model. *Neuroimage* 22:503–520.

38. Kiebel S, Friston K (2004) Statistical parametric mapping for event-related potentials: I. Generic considerations. *Neuroimage* 22:492–502. A

39. Ritter P, Villringer A (2006) Simultaneous EEG-fMRI. *Neurosci Biobehav Rev* 30:823–838.

40. Gotman J, Kobayashi E, Bagshaw AP, Bénar C, Dubeau F (2006) Combining EEG and fMRI: A multimodal tool for epilepsy research. *Journal of Magnetic Resonance Imaging* 23:906–920.

41. Gotman J (2008) Epileptic networks studied with EEG-fMRI. *Epilepsia* 49 Suppl 3:42–51.

42. Gotman J, Pittau F (2011) Combining EEG and fMRI in the study of epileptic discharges. *Epilepsia* 52:38–42.

43. Mulert C et al. (2002) Simultaneous ERP and event-related fMRI: focus on the time course of brain activity in target detection. *Methods Find Exp Clin Pharmacol* 24 Suppl D:17–20.

44. Massimini M et al. (2005) Breakdown of cortical effective connectivity during sleep. *Science* 309:2228–2232.

45. Lioumis P, Kicić D, Savolainen P, Mäkelä JP, Kähkönen S (2009) Reproducibility of TMS-Evoked EEG responses. *Hum Brain Mapp* 30:1387–1396.

46. Miniussi C, Thut G (2010) Combining TMS and EEG offers new prospects in cognitive neuroscience. *Brain Topogr* 22:249–256.

47. Casali AG, Casarotto S, Rosanova M, Mariotti M, Massimini M (2010) General indices to characterize the electrical response of the cerebral cortex to TMS. *Neuroimage* 49:1459–1468.

48. Casarotto S et al. (2010) EEG responses to TMS are sensitive to changes in the perturbation parameters and repeatable over time. *PLoS ONE* 5:e10281.

49. Worsley KJ et al. (1996) A unified statistical approach for determining significant signals in images of cerebral activation. *Hum Brain Mapp* 4:58–73.

50. Grossheinrich N et al. (2009) Theta burst stimulation of the prefrontal cortex: safety and impact on cognition, mood, and resting electroencephalogram. *Biol. Psychiatry* 65:778–784.

51. Litvak V et al. (2011) EEG and MEG data analysis in SPM8. *Comput Intell Neurosci* 2011:852961.

52. Henson RN, Flandin G, Friston KJ, Mattout J (2010) A parametric empirical Bayesian framework for fMRI-constrained MEG/EEG source reconstruction. *Hum Brain Mapp* 31:1512–1531.

53. Kilner JM (2010) Topological inference for EEG and MEG. *Ann. Appl. Stat.* 4:1272–1290.

54. Daunizeau J, Kiebel SJ, Friston KJ (2009) Dynamic causal modelling of distributed electromagnetic responses. *Neuroimage* 47:590–601.

55. Kiebel SJ, Garrido MI, Moran R, Chen C-C, Friston KJ (2009) Dynamic causal modeling for EEG and MEG. *Hum Brain Mapp* 30:1866–1876.

56. Bunzeck N, Doeller CF, Fuentemilla L, Dolan RJ, Duzel E (2009) Reward motivation accelerates the onset of neural novelty signals in humans to 85 milliseconds. *Curr. Biol.* 19:1294–1300.

57. Myatchin I, Mennes M, Wouters H, Stiers P, Lagae L (2009) Working memory in children with epilepsy: an event-related potentials study. *Epilepsy Res.* 86:183–190.

58. Myatchin I, Lagae L (2011) Impaired spatial working memory in children with well-controlled epilepsy: an event-related potentials study. *Seizure* 20:143–150.

59. Rotshtein P et al. (2010) Amygdala damage affects event-related potentials for fearful faces at specific time windows. *Hum Brain Mapp* 31:1089–1105.

60. Whelan R et al. (2010) A high-density ERP study reveals latency, amplitude, and topographical differences in multiple sclerosis patients versus controls. *Clin Neurophysiol* 121:1420–1426.

61. Miller BT, Deouell LY, Dam C, Knight RT, D'Esposito M (2008) Spatio-temporal dynamics of neural mechanisms underlying component operations in working memory. *Brain Res.* 1206:61–75.

62. Holmes AP, Blair RC, Watson JD, Ford I (1996) Nonparametric analysis of statistic images from functional mapping experiments. *J. Cereb. Blood Flow Metab.* 16:7–22.

63. Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp* 15:1–25.

64. Holmes AP, Blair RC, Watson JD, Ford I (1996) Nonparametric analysis of statistic images from functional mapping experiments. *J. Cereb. Blood Flow Metab* 16:7–22.

65. Blair RC, Karniski W (1993) An alternative method for significance testing of waveform difference potentials. *Psychophysiology* 30:518–524.

66. Huber R et al. (2007) TMS-induced cortical potentiation during wakefulness locally increases slow wave activity during sleep. *PLoS ONE* 2:e276.

67. Groppe DM et al. (2010) The phonemic restoration effect reveals pre-N400 effect of supportive sentence context in speech perception. *Brain Res.* 1361:54–66.

68. Bobes MA, Quiñonez I, Perez J, Leon I, Valdés-Sosa M (2007) Brain potentials reflect access to visual and emotional memories for faces. *Biol Psychol* 75:146–153.

69. Groppe DM, Urbach TP, Kutas M (2011) Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. *Psychophysiology* 48:1711–1725.

70. Bekinschtein TA et al. (2009) Neural signature of the conscious processing of auditory regularities. *Proc. Natl. Acad. Sci. U.S.A* 106:1672–1677.

71. Kissler J, Koessler S (2011) Emotionally positive stimuli facilitate lexical decisions--An ERP study. *Biological Psychology* 86:254–264.

72. Sergent C, Baillet S, Dehaene S (2005) Timing of the brain events underlying access to consciousness during the attentional blink. *Nat. Neurosci* 8:1391–1400.

73. Poline JB, Mazoyer BM (1994) Analysis of individual brain activation maps using hierarchical description and multiscale detection. *IEEE Trans Med Imaging* 13:702–710.

74. Maris E (2004) Randomization tests for ERP topographies and whole spatiotemporal data matrices. *Psychophysiology* 41:142–151.

75. Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* 164:177–190.

76. Bullmore ET et al. (1999) Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans Med Imaging* 18:32–42.

77. Tillikainen L, Salli E, Korvenoja A, Aronen HJ (2006) A cluster mass permutation test with contextual enhancement for fMRI activation detection. *Neuroimage* 32:654–664.

78. Hayasaka S, Nichols TE (2003) Validating cluster size inference: random field and permutation methods. *Neuroimage* 20:2343–2356.

79. Hayasaka S, Nichols TE (2004) Combining voxel intensity and cluster extent with permutation test framework. *Neuroimage* 23:54–63.

80. Hayasaka S, Phan KL, Liberzon I, Worsley KJ, Nichols TE (2004) Nonstationary cluster-size inference with random field and permutation methods. *Neuroimage* 22:676–687.

81. Smith SM, Nichols TE (2009) Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage* 44:83–98.

82. Salimi-Khorshidi G, Smith SM, Nichols TE (2011) Adjusting the effect of nonstationarity in cluster-based and TFCE inference. *Neuroimage* 54:2006–2019.

83. Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134:9–21.

84. He H, Garcia E (2009) Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* 21:1263–1284.

85. Davis J, Goadrich M (2006) in *Proceedings of the 23rd international conference on Machine learning*, ICML '06. (ACM, New York, NY, USA), pp 233–240.

86. Brodersen KH, Ong CS, Stephan KE, Buhmann JM (2010) in *2010 20th International Conference on Pattern Recognition (ICPR)* (IEEE), pp 4263–4266.

87. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405:442–451.

88. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16:412–424.

89. Lage-Castellanos, Martínez-Montes E, Hernández-Cabrera JA, Galán L (2010) False discovery rate and permutation test: An evaluation in ERP data analysis. *Statistics in Medicine* 29:63–74.

90. Groppe DM, Urbach TP, Kutas M (2011) Mass univariate analysis of event-related brain potentials/fields II: Simulation studies. *Psychophysiology* 48:1726–1737.

91. Delorme A, Makeig S, Fabre-Thorpe M, Sejnowski T (2002) From single-trial EEG to brain area dynamics. *Neurocomputing* 44-46:1057–1064.

92. Delorme A, Rousselet G, M Macé, Fabre-Thorpe M (2004) Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cognitive Brain Research* 19:103–113.

93. Manly BFJ (1997) *Randomization, bootstrap and Monte Carlo methods in biology* (Chapman & Hall/ CRC).

94. Suckling J, Bullmore E (2004) Permutation tests for factorially designed neuroimaging experiments. *Hum Brain Mapp* 22:193–205.

95. Anderson MJ, Legendre P (1999) An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation* 62:271–303.

96. Anderson MJ, Robinson J (2001) Permutation Tests for Linear Models. *Australian & New Zealand Journal of Statistics* 43:75–88.

97. Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26:32–46.

98. Anderson M, Braak CT (2003) Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation* 73:85–113.

99. Posner MI (1980) Orienting of attention. *Quarterly Journal of Experimental Psychology* 32:3–25.

100. Posner M, Cohen Y (1984) Components of visual orienting. *Attention and Performance, X, Hillsdale*.

101. Posner MI, Walker JA, Friedrich FJ, Rafal RD (1984) Effects of parietal injury on covert orienting of attention. *J. Neurosci.* 4:1863–1874.

102. Posner MI, Rafal RD, Choate LS, Vaughan J (1985) Inhibition of return: Neural basis and function. *Cognitive Neuropsychology* 2:211–228.

103. Lupiáñez J et al. (2004) Independent effects of endogenous and exogenous spatial cueing: inhibition of return at endogenously attended target locations. *Exp Brain Res* 159:447–457.

104. Lupianez J, Klein RM, Bartolomeo P (2006) Inhibition of return: Twenty years after. *Cogn Neuropsychol* 23:1003–1014.

105. Prime DJ, Ward LM (2006) Cortical expressions of inhibition of return. *Brain Res.* 1072:161–174.

106. Tian Y, Yao D (2008) A study on the neural mechanism of inhibition of return by the event-related potential in the Go/NoGo task. *Biol Psychol* 79:171–178.

107. Prime DJ, Jolicoeur P (2009) Response-selection conflict contributes to inhibition of return. *J Cogn Neurosci* 21:991–999.

108. Prime DJ, Jolicoeur P (2009) On the relationship between occipital cortex activity and inhibition of return. *Psychophysiology* 46:1278–1287.

109. Tian Y, Klein RM, Satel J, Xu P, Yao D (2011) Electrophysiological explorations of the cause and effect of inhibition of return in a cue-target paradigm. *Brain Topogr* 24:164–182.

110. Brainard DH (1997) The Psychophysics Toolbox. *Spat Vis* 10:433–436.

111. Pourtois G, Delplanque S, Michel C, Vuilleumier P (2008) Beyond conventional event-related brain potential (ERP): exploring the time-course of visual emotion processing using topographic and principal component analyses. *Brain Topogr* 20:265–277.

112. Lehmann D, Skrandies W (1980) Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalogr Clin Neurophysiol* 48:609–621.

113. Worsley KJ, Andermann M, Koulis T, MacDonald D, Evans AC (1999) Detecting changes in nonisotropic images. *Hum Brain Mapp* 8:98–101.

114. Oostenveld R, Fries P, Maris E, Schoffelen J-M (2011) FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci* 2011:156869. A

115. Tadel F, Baillet S, Mosher JC, Pantazis D, Leahy RM (2011) Brainstorm: A User-Friendly Application for MEG/EEG Analysis. *Comput Intell Neurosci* 2011:879716.

116. Dalal SS et al. (2011) MEG/EEG source reconstruction, statistical evaluation, and visualization with NUTMEG. *Comput Intell Neurosci* 2011:758973.

117. Clemens B et al. (2008) Three-dimensional localization of abnormal EEG activity in migraine: a low resolution electromagnetic tomography (LORETA) study of migraine patients in the pain-free interval. *Brain Topogr* 21:36–42.

118. Clemens B et al. (2011) EEG functional connectivity of the intrahemispheric cortico-cortical network of idiopathic generalized epilepsy. *Epilepsy Res*.

119. Wendel K et al. (2009) EEG/MEG source imaging: methods, challenges, and open issues. *Comput Intell Neurosci*:656092.